Variation in actual relationship as a consequence of Mendelian sampling and linkage

Genet. Res., Camb. (2011), **93**, *pp.* 47–64. © Cambridge University Press 2011
doi:10.1017/S0016672310000480

Variation in actual relationship as a consequence of Mendelian sampling and linkage

W.G. HILL[1]* AND B.S. WEIR[2]
[1] *Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, UK*
[2] *Department of Biostatistics, University of Washington, Box 357232, Seattle, WA 98195-7232, USA*

(*Received 3 June 2010 and in revised form 30 September 2010; first published online 12 January 2011*)

Summary

Although the expected relationship or proportion of genome shared by pairs of relatives can be obtained from their pedigrees, the actual quantities deviate as a consequence of Mendelian sampling and depend on the number of chromosomes and map length. Formulae have been published previously for the variance of actual relationship for a number of specific types of relatives but no general formula for non-inbred individuals is available. We provide here a unified framework that enables the variances for distant relatives to be easily computed, showing, for example, how the variance of sharing for great grandparent–great grandchild, great uncle–great nephew, half uncle–nephew and first cousins differ, even though they have the same expected relationship. Results are extended in order to include differences in map length between sexes, no recombination in males and sex linkage. We derive the magnitude of skew in the proportion shared, showing the skew becomes increasingly large the more distant the relationship. The results obtained for variation in actual relationship apply directly to the variation in actual inbreeding as both are functions of genomic coancestry, and we show how to partition the variation in actual inbreeding between and within families. Although the variance of actual relationship falls as individuals become more distant, its coefficient of variation rises, and so, exacerbated by the skewness, it becomes increasingly difficult to distinguish different pedigree relationships from the actual fraction of the genome shared.

1. Introduction

Characterizing the relationship between pairs of individuals continues to be of importance in many areas of population and quantitative genetics. Variation in genome sharing identical by descent (ibd) over the genome depends both on the pedigree and the extent to which alleles at different loci are jointly ibd. The degree of relationship might be inferred from pedigree information or it can be estimated from genetic information (Weir *et al*., 2006; Visscher *et al*., 2006; Yu *et al*., 2006), but in either case there is variation in relationship measures. A recent development has been to utilize this variability in the actual relationship to estimate the components of variance for quantitative traits from the variation in resemblance among full sibs, i.e. family members who have the same pedigree relationship (Visscher *et al*., 2006).

By making assumptions about the mapping function, the variation in the proportion of genome-shared ibd, or actual relationship, can be computed for different pedigrees. Formulae have been published for autosomal loci of lineal descendants (Stam & Zeven, 1981; Hill, 1993*a*), sibs (Hill, 1993*b*) and other relatives, including cousins (Guo, 1995). Formulae have also been given for the variation of identity of full sibs for both alleles at each site (Visscher *et al*., 2006) and for sex-linked loci (Visscher, 2009).

These analyses are solely concerned with the variances of the distributions of sharing. The distribution itself or other functions of it have also been obtained. In particular, Donnelly (1983) computed the probability that the proportion shared with an ancestor exceeded zero. Bickeboller & Thompson (1996*a*, *b*)

* Corresponding author. Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, UK. Tel: +44-(0)131-650 5705. Fax: +44-(0)131-650 6564. e-mail: w.g.hill@ed.ac.uk

obtained approximations for the distribution of the proportion shared between half-sibs and between offspring and parent. The full distribution has been obtained by Stefanov and colleagues for lineal descendants (Stefanov, 2000, 2004) and for half sibs (Ball & Stevanov, 2005). Their results generally take the form of a set of equations and computer routines for numerical evaluation.

With the advent of dense genome mapping, it has become possible to estimate the actual proportion of the genome shared for pairs of relatives and to compare the observed with expected values. This has been done for full sibs by Visscher *et al*. (2006, 2007), and there was generally good agreement between observed and expected sharing.

Mapping with multiple markers enables relatives to be identified among samples from the population. The ability to correctly assign relationship, to distinguish between second and third cousins, for example, depends on the sampling variance of the actual proportion of genome shared and the additional sampling due to the use of a limited number of markers. Such data arise in genome-wide association studies, for example, where up to millions of single nucleotide polymorphism (SNP) markers are genotyped on thousands of individuals, and the relationship structure of the data is an important component in determining the reliability of conclusions on trait gene identification. Genetic variances of quantitative traits can be estimated by taking advantage of the variation in genome sharing to account for phenotypic similarity both within families of full sibs (including dizygotic twins) (Visscher *et al*., 2006, 2007) and between families utilizing information on distant relatives not available from known relationships (Yang *et al*., 2010). Quantifying the degree of relationship is also an important aspect of genotype data cleaning in genome-wide association studies (Laurie *et al*., 2010), for guarding against incorrect annotation of family membership or for modifying tests of marker trait association (Choi *et al*., 2009). Genomic selection, which utilizes dense mapping for identifying sharing of genes among relatives, depends on there being variability in genome sharing of relatives that have the same pedigree relationship (Meuwissen *et al*., 2001), and which has major application, mainly so far in plant and animal breeding. It may be based directly on the actual genomic relationship matrix or with weighting dependent on the variance in the trait associated with particular genomic regions (Goddard, 2009). These activities require an appreciation of the extent of the variation in genome sharing by identity and have motivated this study.

Our objective in this paper is to consider moments of the distribution of allele sharing, and to obtain formulae that can be applied simply to any kind and degree of relationship, including direct descendants and those of half- and of full sibs. The distributions can be highly skewed, particularly when the relationship is low, and hence we also obtain formulae for the magnitude of skew of relationship. Although we restrict the analysis to the relationship among non-inbred individuals, the results apply directly to the variation in actual inbreeding of offspring of consanguineous matings and we show how to apply them.

2. General formulae for variance of genome sharing of non-inbred individuals

(i) *Background theory*

At any locus individuals may share zero, one or two pairs of alleles ibd with probabilities k_0, k_1 or k_2. The actual ibd status can be indicated by $\breve{k}_m$, $m=0, 1, 2$, where $\breve{k}_m=1$ if the individuals share exactly m pairs of alleles ibd and $\breve{k}_m=0$ otherwise. The probabilities k_m depend on the pedigree structure and are the expected values of the $\breve{k}_m$. As exactly one of the $\breve{k}_m$ is equal to 1 at any locus and as squaring an indicator does not change its value, their variances and covariances are

$$\mu_2(\breve{k}_m)=\mathrm{Var}(\breve{k}_m)=k_m(1-k_m), \quad m=0,\ 1,\ 2,$$
$$\mathrm{Cov}(\breve{k}_m,\breve{k}_{m'})=-k_m k_{m'}, \quad m\neq m'.$$

Less detailed measures of relationship are the coancestry or kinship coefficient, $\theta=\frac{1}{2}k_2+\frac{1}{4}k_1$, the probability that an allele drawn at random from one individual is ibd to a random allele from the other, and the relationship $R=2\theta=k_2+\frac{1}{2}k_1$. This equals Wright's (1922) relationship for non-inbred individuals and is also called the 'numerator relationship'. We shall primarily use R here as we are considering an analysis of genome sharing, for R is the probability that a random allele identified in one individual is present ibd in the other. We have previously considered variation in actual coancestry (Cockerham & Weir, 1983; Weir *et al*., 2005) and thus in relationship. The actual relationship is $\breve{R}=\breve{k}_2+\frac{1}{2}\breve{k}_1$ and this has variance

$$\mu_2(\breve{R})=\mathrm{Var}(\breve{R})=k_2+\tfrac{1}{4}k_1-\left(k_2+\tfrac{1}{2}k_1\right)^2=k_2+\tfrac{1}{4}k_1-R^2. \quad (1)$$

The quantity $\frac{1}{4}k_2+\frac{1}{16}k_1$ was written as Δ by Cockerham & Weir (1983) and is the probability that two pairs of alleles at the same locus are ibd.

The inbreeding coefficient F is the probability that the two alleles carried by an individual are ibd. We have discussed the variation in actual inbreeding (Weir *et al*., 1980; Cockerham & Weir, 1983), with the variation in the two-allele measures θ and F expressed as a function of the ibd probability of a set of two, three or four alleles. We shall also discuss coefficients of variation of actual identity. For example,

$$\mathrm{CV}(\breve{R})=\sqrt{\mathrm{Var}(\breve{R})}/E(\breve{R})=\sqrt{\mathrm{Var}(\breve{R})}/R.$$

Table 1. *Expectations and variances for actual identity at individual loci*

Relationship	R	θ	k_0	k_1	k_2	$\mathrm{Var}(\breve{k}_1)$	$\mathrm{SD}(\breve{R})$	$\mathrm{CV}(\breve{R})$
Parent–offspring	0·5	0·25	0	1	0	0	0	0
Full sibs	0·5	0·25	0·25	0·5	0·25	0·25	0·3536	0·707
Grandparent–grandoffspring	0·25	0·125	0·5	0·5	0·0	0·25	0·2500	1·000
Half sibs	0·25	0·125	0·5	0·5	0·0	0·25	0·2500	1·000
Uncle–nephew	0·25	0·125	0·5	0·5	0·0	0·25	0·2500	1·000
Double first cousins	0·25	0·125	0·5625	0·375	0·0625	0·2344	0·3062	1·225
Greatgrandparent–greatgrandoffspring	0·125	0·0625	0·75	0·25	0·0	0·1875	0·2165	1·732
Half uncle–nephew	0·125	0·0625	0·75	0·25	0·0	0·1875	0·2165	1·732
First cousins	0·125	0·0625	0·75	0·25	0·0	0·1875	0·2165	1·732
Great uncle–great nephew	0·125	0·0625	0·75	0·25	0·0	0·1875	0·2165	1·732
Half cousins	0·0625	0·0313	0·875	0·125	0·0	0·1094	0·1654	2·645
Cousins once removed	0·0625	0·0313	0·875	0·125	0·0	0·1094	0·1654	2·645
Second cousins	0·0313	0·0156	0·9375	0·0625	0·0	0·0586	0·1210	3·872

In Table 1, we list values for the ks, R and their single-locus variances and covariances for some common relationships. We now consider the variances and covariances of the actual identities when that they are averaged over the genome, assuming that they have the same expected values at all loci. The results for single loci also apply if the loci are completely linked and are therefore a limiting case of the genome-average results.

When we consider the variation in sharing of relatives over the genome, we require the average over pairs of loci i, j of the covariances of the actual sharing indicators $\breve{k}_i, \breve{k}_j$ for 0, 1 or 2 pairs of alleles. For a set of r loci $\breve{k} = \frac{1}{r}\sum_{i=1}^{r}\breve{k}_i$ and

$$E(\breve{k}^2) = \frac{1}{r^2}E\left(\sum_i \breve{k}_i^2 + \sum_{i\neq j}\breve{k}_i\breve{k}_j\right).$$

Combining the two terms in this sum and subtracting the square of the mean gives

$$\mathrm{Var}(\breve{k}) = \frac{1}{r^2}E\left(\sum_i\sum_j \breve{k}_i\breve{k}_j\right) - k^2$$

and similar arguments apply to higher moments discussed later.

(ii) *Lineal descendants*

If g generations separate two individuals, one being a lineal descendant of the other, $k_2 = 0$ and $k_1 = (\frac{1}{2})^{g-1}$. For a parent and offspring pair ($g = 1$, e.g. **A** and **D** in Fig. 1), $\breve{k}_1 = k_1 = 1$ and $\mathrm{Var}(\breve{k}_1) = 0$. For linked gametic loci i, j the only way both values can be equal to one in subsequent generations (e.g. **G**, **J**) is if there has been no recombination in the descent from ancestor to descendant. The expected value of their product is therefore

$$E(\breve{k}_{1i}\breve{k}_{1j}) = \left(\tfrac{1}{2}(1-c_{ij})\right)^{g-1},$$

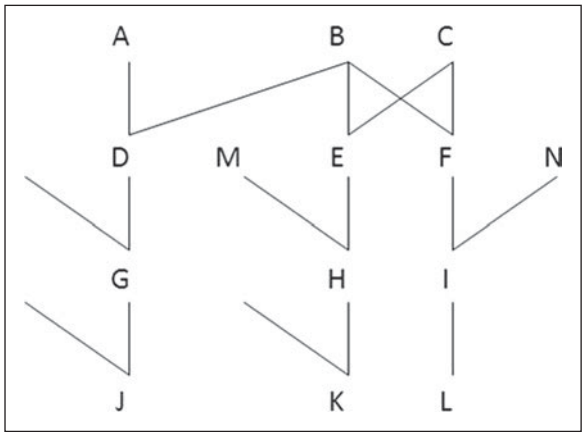

Fig. 1. Examples of relationship

Relationship	*Example*	*R*	*Relationship*	*Example*	*R*
Lineal relatives			*Full sibs and their descendants*		
Parent–offspring	**AD**	$\frac{1}{2}$	Full sibs	**EF**	$\frac{1}{2}$
Grandparent–grandoffspring	**AG**	$\frac{1}{4}$	Uncle–nephew	**EI**	$\frac{1}{4}$
Greatgrandparent–ggoffspring	**AJ**	$\frac{1}{8}$	Great uncle–gt nephew	**EL**	$\frac{1}{8}$
Half-sibs and their descendants			(First) cousins	**HI**	$\frac{1}{8}$
Half-sibs	**DE**	$\frac{1}{4}$	Cousins once removed	**HL**	$\frac{1}{16}$
Half-uncle–nephew	**DH**	$\frac{1}{8}$	Second cousins	**KL**	$\frac{1}{32}$
Half-cousins	**GH**	$\frac{1}{16}$	Double first cousins*	**HI**	$\frac{1}{4}$

*If **M** and **N** are also full sibs.

where c_{ij} is the recombination fraction between loci. For convenience, we will drop the ij subscript on c_{ij}. The covariance of these two variables is

$$\mathrm{Cov}(\breve{k}_{1i}, \breve{k}_{1j}) = \left(\tfrac{1}{2}(1-c)\right)^{g-1} - \left(\tfrac{1}{4}\right)^{g-1}. \quad (2)$$

Note that this covariance is zero if the loci are unlinked and $c = 0{\cdot}5$, or if one individual is the offspring

of the other and $g=1$. Setting $c=0$ gives the variance $k_{1i}(1-k_{1i})$ as the two loci are then transmitted as a unit.

For allele sharing over the whole genome, suppose there are infinitely many loci along a chromosome of length l and further suppose Haldane's (1919) mapping function holds so that $(1-c)=\frac{1}{2}(1+\mathrm{e}^{-2d})$, where d is the map length between loci i, j. Therefore, from eqn (2),

$$\mathrm{Cov}(\breve{k}_{1i},\breve{k}_{1j})=\left(\tfrac{1}{4}\right)^{g-1}\left[(1+\mathrm{e}^{-2d})^{g-1}-1\right].$$

The variance of allele sharing over the whole chromosome is the average of all the covariances and this can be calculated as an integral by letting x, y be the positions of pairs of loci:

$$\mathrm{Var}_{\mathrm{Lin},g}(\breve{k}_1)= \frac{2}{l^2}\left(\frac{1}{4}\right)^{g-1}\int_{x=0}^{l}\int_{y=0}^{x}\left[(1+\mathrm{e}^{-2(x-y)})^{g-1}-1\right]\mathrm{d}y\,\mathrm{d}x \tag{3}$$

(Stam & Zeven, 1981; Hill, 1993*a*).

As we use this function repeatedly and more generally subsequently, we define

$$\phi_n(l)=\frac{2}{l^2}\left(\frac{1}{4}\right)^{n}\int_{x=0}^{l}\int_{y=0}^{x}\left[(1+\mathrm{e}^{-2(x-y)})^{n}-1\right]\mathrm{d}y\;\mathrm{d}x \tag{4a}$$

$$=\begin{cases}\dfrac{1}{2l^2}\left(\dfrac{1}{4}\right)^{n}\displaystyle\sum_{r=1}^{n}\binom{n}{r}\left[\frac{2rl-1+\mathrm{e}^{-2rl}}{r^2}\right], & n\geqslant 1,\\ 0, & n=0.\end{cases} \tag{4b}$$

(Hill, 1993*a*). At the limits, for $l\to 0$, $\phi_n(l)\to(\frac{1}{2})^n\times[1-(\frac{1}{2})^n]$ and for $l\to\infty$, $\phi_n(l)\to 0$. The variance of the chromosome-sharing variable $\breve{k}_1$ for lineal relatives g generations apart can then be expressed as $\mathrm{Var}_{\mathrm{Lin},g}(\breve{k}_1,l)=\phi_{g-1}(l)$. Also $\mathrm{Var}_{\mathrm{Lin},g}(\breve{R},l)=\frac{1}{4}\mathrm{Var}_{\mathrm{Lin},g}(\breve{k}_1,l)$ and $\mathrm{Var}_{\mathrm{Lin},g}(\breve{\theta},l)=\frac{1}{16}\mathrm{Var}_{\mathrm{Lin},g}(\breve{k}_1,l)$. The coefficient of variation (CV) of $\breve{k}_1$ is given by

$$\mathrm{CV}_{\mathrm{Lin},g}(\breve{k}_1,l)=2^{g-1}\sqrt{\phi_{g-1}(l)} =\frac{1}{l}\left\{\frac{1}{2}\sum_{r=1}^{g-1}\binom{g-1}{r}\left[\frac{2rl-1+\mathrm{e}^{-2rl}}{r^2}\right]\right\}^{1/2},\quad g\geqslant 2$$

(Visscher, 2009) and is the same for $\breve{R}$ and $\breve{\theta}$. For a whole genome comprising K chromosomes of lengths $l_1, l_2, \ldots, l_K$ and total map length $L=\sum_{i=1}^{K}l_i$, the variance is

$$\mathrm{Var}_{\mathrm{Lin},g}(\breve{k}_1)=\frac{1}{L^2}\sum_{i=1}^{K}l_i^2\phi_{g-1}(l_i). \tag{5}$$

We now evaluate the variance of genome sharing or relationship among collateral relatives and their descendants using eqns (3) and (4). Results are summarized in Box 1.

(iii) *Half-sibs and their descendants*

(a) *General formulation*

Just as for lineal relatives, half-sibs (e.g. **D** and **E** in Fig. 1) and their descendants can have only one or zero pairs of ibd alleles at a locus. Formulae for variances of sharing ibd for half-sibs were given by Hill (1993*b*) and Guo (1995), but we generalize these here in order to include subsequent generations.

The probability that half-sibs share one pair of alleles is $E(\breve{k}_1)=k_1=\frac{1}{2}$ and the probability that they share zero pairs is $\breve{k}_0=\frac{1}{2}$, so $\mathrm{Var}(\breve{k}_1)=\frac{1}{4}$. Half-sibs share one pair of alleles at each of loci i, j only if they both receive the same non-recombinant or the same recombinant haplotype from their common parent. Therefore,

$$E(\breve{k}_{1i}\breve{k}_{1j})=\tfrac{1}{2}(1-c)^2+\tfrac{1}{2}c^2 \tag{6}$$

and the covariance of the allele-sharing indicators is

$$\mathrm{Cov}(\breve{k}_{1i},\breve{k}_{1j})=\tfrac{1}{2}(1-c)^2+\tfrac{1}{2}c^2-\tfrac{1}{4}=\tfrac{1}{4}(1-2c)^2 \tag{7}$$

showing that the covariance of $\breve{k}$s for unlinked loci is zero.

When we consider relationships across generations, for example, half-uncle nephew, the probability that these share haplotypes is proportional to $\frac{1}{2}(1-c)$ of the probability that the half-sibs share haplotypes. For half-sibs and other relatives who are not lineal descendents, the probability of sharing is not simply proportional to powers of $(1-c)$ but involve others such as c^2 as shown in eqn (6). In order to generalize formulae across generations, we find it convenient to express all powers of c in terms of $b=\frac{1}{2}(1-c)$ as

$$c^n=\left[1-2\left(\frac{1}{2}(1-c)\right)\right]^n=\sum_{i=0}^{n}\binom{n}{i}(-2b)^i. \tag{8}$$

Therefore, from eqn (6), for half-sibs

$$E(\breve{k}_{1i}\breve{k}_{1j})=4b^2-2b+\tfrac{1}{2}.$$

This is a specific example of expressions which appear in all succeeding analyses, and so we consider the general form

$$E(\breve{k}_{1i}\breve{k}_{1j})=\sum_n a_n b^n. \tag{9}$$

For unlinked loci, $b=\frac{1}{4}$, the $\breve{k}$s are independent and (9) gives the product of the expected values of $\breve{k}_{1i}$ and $\breve{k}_{1j}$, so

$$\mathrm{Cov}(\breve{k}_{1i},\breve{k}_{1j})=\sum_n a_n\left\lfloor b^n-\left(\tfrac{1}{4}\right)^n\right\rfloor.$$

Expressed in terms of map positions x, y for these loci, $b=\frac{1}{4}(1+\mathrm{e}^{-2(x-y)})$ and

$$\mathrm{Cov}(\breve{k}_{1i},\breve{k}_{1j})=\sum_n a_n\left(\tfrac{1}{4}\right)^n\left\lfloor\left(1+\mathrm{e}^{-2(x-y)}\right)^n-1\right\rfloor.$$

Box 1. *Summary of formulae for variances of genome sharing.* $R=\left(\frac{1}{2}\right)^g$

A. Unilineal relatives ($k_2=0$ and $\mathrm{Var}(\breve{R},l)=\frac{1}{4}\mathrm{Var}(\breve{k}_1,l)$)

Lineal descendants

$\mathrm{Var}_{\mathrm{Lin},g}(\breve{k}_1,l)=\phi_{g-1}(l)$.

Examples: $g=1$ for parent–offspring (when $\mathrm{Var}_{\mathrm{Lin},g}(\breve{k}_1,l)=0$), $g=2$ for grandparent–grandoffspring.

Half-sibs and their descendants

$\mathrm{Var}_{\mathrm{HS},g}(\breve{k}_1,l)=4\phi_g(l)-2\phi_{g-1}(l)+\frac{1}{2}\phi_{g-2}(l)$.

Examples: $g=2$ for half sibs, $g=3$ for half uncle-nephew, $g=4$ for half cousins.

Descendants of full sibs

Uncle–nephew and nephew's descendants

$\mathrm{Var}_{\mathrm{UN},g}(\breve{k}_1,l)=8\phi_{g+1}(l)-4\phi_g(l)+\frac{1}{2}\phi_{g-1}(l)+\frac{1}{4}\phi_{g-2}(l)$.

Examples: $g=2$ for uncle-nephew, $g=3$ for great uncle-great nephew.

Cousins and descendants

$\mathrm{Var}_{\mathrm{C},g}(\breve{k}_1,l)=8\phi_{g+1}(l)-4\phi_g(l)+\frac{3}{2}\phi_{g-1}(l)-\frac{1}{2}\phi_{g-2}(l)+\frac{1}{8}\phi_{g-3}(l)$.

Examples: $g=3$ for (first) cousins, $g=5$ for second cousins or cousins twice removed.

B. Bilineal relatives ($k_2\neq 0$)

Full sibs

$\mathrm{Var}_{\mathrm{FS}}(\breve{R},l)=2\phi_2(l)-\phi_1(l)$.

$\mathrm{Var}_{\mathrm{FS}}(\breve{k}_2,l)=\mathrm{Var}_{\mathrm{FS}}(\breve{k}_0,l)=16\phi_4(l)-16\phi_3(l)+8\phi_2(l)-2\phi_1(l)$,

$\mathrm{Var}_{\mathrm{FS}}(\breve{k}_1,l)=4\mathrm{Var}_{\mathrm{FS}}(\breve{k}_2,l)-4\mathrm{Var}_{\mathrm{FS}}(\breve{R},l)$,

$\mathrm{Cov}_{\mathrm{FS}}(\breve{k}_2,\breve{k}_1,l)=\mathrm{Cov}_{\mathrm{FS}}(\breve{k}_1,\breve{k}_0,l)=-2\mathrm{Var}_{\mathrm{FS}}(\breve{k}_2,l)+2\mathrm{Var}_{\mathrm{FS}}(\breve{R},l)$,

$\mathrm{Cov}_{\mathrm{FS}}(\breve{k}_2,\breve{k}_0,l)=\mathrm{Var}_{\mathrm{FS}}(\breve{k}_2,l)-2\mathrm{Var}_{\mathrm{FS}}(\breve{R},l)$.

Double first cousins

$\mathrm{Var}_{\mathrm{DFC}}(\breve{R},l)=4\phi_4(l)-2\phi_3(l)+\frac{3}{4}\phi_2(l)-\frac{1}{4}\phi_1(l)$,

$\mathrm{Var}_{\mathrm{DFC}}(\breve{k}_2,l)=64\phi_8(l)-64\phi_7(l)+40\phi_6(l)-20\phi_5(l)+\frac{33}{4}\phi_4(l)-\frac{5}{2}\phi_3(l)+\frac{5}{8}\phi_2(l)-\frac{1}{8}\phi_1(l)$,

$\mathrm{Var}_{\mathrm{DFC}}(\breve{k}_1,l)=4\mathrm{Var}_{\mathrm{DFC}}(\breve{k}_2,l)$,

$\mathrm{Var}_{\mathrm{DFC}}(\breve{k}_0,l)=\mathrm{Var}_{\mathrm{DFC}}(\breve{k}_2,l)+2\mathrm{Var}_{\mathrm{DFC}}(\breve{R},l)$,

$\mathrm{Cov}_{\mathrm{DFC}}(\breve{k}_2,\breve{k}_1,l)=-2\mathrm{Var}_{\mathrm{DFC}}(\breve{k}_2,l)+\mathrm{Var}_{\mathrm{DFC}}(\breve{R},l)$,

$\mathrm{Cov}(\breve{k}_2,\breve{k}_0,l)=\mathrm{Var}_{\mathrm{DFC}}(\breve{k}_2,l)-\mathrm{Var}_{\mathrm{DFC}}(\breve{R},l)$,

$\mathrm{Cov}(\breve{k}_1,\breve{k}_0,l)=-2\mathrm{Var}_{\mathrm{DFC}}(\breve{k}_2,l)-\mathrm{Var}_{\mathrm{DFC}}(\breve{R},l)$.

Using eqns (3) and (4), we obtain

$$\mathrm{Var}(\breve{k}_1,l)=\sum_n a_n\phi_n(l). \tag{10}$$

Applying this methodology to half-sibs,

$\mathrm{Var}_{\mathrm{HS}}(\breve{k}_1,l)=4\phi_2(l)-2\phi_1(l)+\frac{1}{2}\phi_0(l)=4\phi_2(l)-2\phi_1(l)$,

because $\phi_0(l)=0$. Also

$$\mathrm{Var}_{\mathrm{HS}}(\breve{R},l)=\phi_2(l)-\tfrac{1}{2}\phi_1(l). \tag{11}$$

(b) *Half-uncle nephew and descendants*

The probability that half-uncle and nephew (e.g. **D** and **H** in Fig. 1; or, implicit here and subsequently,

half-aunt and nephew or niece, etc.) share one pair of alleles ibd is $k_1=\frac{1}{4}$. They share a pair of alleles ibd at loci i and j only if **H** receives from its parent **E** the non-recombinant haplotype that carries alleles from **B**, the common parent of **D** and **E**. Therefore

$$E(\breve{k}_{1i}\breve{k}_{1j})=\tfrac{1}{2}(1-c)\left[\tfrac{1}{2}(1-c)^2+\tfrac{1}{2}c^2\right]=4b^3-2b^2+\tfrac{1}{2}b$$

and immediately, by using (9) and (10),

$$\mathrm{Var}_{\mathrm{HUN}}(\breve{k}_1,l)=4\phi_3(l)-2\phi_2(l)+\tfrac{1}{2}\phi_1(l).$$

We generalize the formulae with reference to pairs of relatives that are g generations apart, i.e. their pedigree relationship is $(\frac{1}{2})^g$. Thus, $g=2$ for half sibs (and grandparent–grandoffspring, as above), $g=3$ for half-uncle nephew and $g=4$ for half-cousins (**G** and **H** in Fig. 1) and for half-great uncle nephew (**D** and **K**). The one-locus allele sharing indicator has expectation $E(\breve{k}_1)=(0{\cdot}5)^{g-1}$ and those for two loci reduce by a proportion $\frac{1}{2}(1-c)$ each generation as the g meioses are independent. Hence

$$\mathrm{Var}_{\mathrm{HS},g}(\breve{k}_1,l)=4\phi_g(l)-2\phi_{g-1}(l)+\tfrac{1}{2}\phi_{g-2}(l).$$

Setting $g=2$ and noting that $\phi_0(l)=0$ provide the half-sib result. Note also that the variances are the same for any collateral and lineal offspring of half-sibs that have the same relationship, e.g. half-cousins and half great uncle–great nephew.

(iv) *Lineal descendants of full-sibs*

We now discuss the relationships between full sibs and their lineal descendants and among these descendants, where it is still the case that only one or zero pairs of alleles might be ibd, i.e. $k_2=0$. We defer to the next section a treatment of full sibs and of bilineal relatives in general where $k_2>0$. Note, however, that since the maternal and paternal transmissions are independent,

$$\mathrm{Var}_{\mathrm{FS}}(\breve{R},l)=2\phi_2(l)-\phi_1(l),$$

i.e. twice that for half-sibs (eqn (11)) (Hill, 1993*b*; Guo, 1995).

(a) *Uncle–nephew*

In Fig. 1, **E** and **F** are full sibs and **I** is the offspring of **F** and a nephew of **E**. At any locus, they can share one or zero pairs of alleles with probabilities $k_1=k_0=\frac{1}{2}$. They can share a pair of alleles ibd at loci i and j in two ways: either **I** receives a non-recombinant haplotype from **F**, and **E**, **F** both carry copies of that haplotype which might themselves be both recombinant or non-recombinant from one of their parents; or **I** receives a recombinant haplotype from **F**, and **E**, **F** receive ibd alleles at i from one parent and ibd alleles at j from the other. So

$$\begin{aligned}E(\breve{k}_{1i}\breve{k}_{1j})&=(1-c)\left[\tfrac{1}{2}(1-c)^2+\tfrac{1}{2}c^2\right]+\tfrac{1}{4}c\\&=8b^3-4b^2+\tfrac{1}{2}b+\tfrac{1}{4}.\end{aligned}\qquad(12)$$

Integrating over a chromosome of length l and using (9) and (10)

$$\begin{aligned}&\mathrm{Var}_{\mathrm{UN}}(\breve{k}_1,l)=8\phi_3(l)-4\phi_2(l)+\tfrac{1}{2}\phi_1(l),\\&\mathrm{Var}_{\mathrm{UN}}(\breve{R},l)=2\phi_3(l)-\phi_2(l)+\tfrac{1}{8}\phi_1(l).\end{aligned}$$

These results are not the same as those for half-sibs, even though the single-locus probabilities k_0, k_1 are the same nor are they twice the value for half-uncle nephew.

(b) *Uncle and descendants of a nephew*

For great-uncle nephew (e.g. **E** and **L** in Fig. 1) and further descendents of the nephew, results are obtained immediately from (12) as the expressions are multiplied by further coefficients b. Hence, if they are g generations apart

$$\mathrm{Var}_{\mathrm{UN},g}(\breve{R},l)=2\phi_{g+1}(l)-\phi_g(l)+\tfrac{1}{8}\phi_{g-1}(l)+\tfrac{1}{16}\phi_{g-2}(l).$$

This reduces to the uncle–nephew case (where $R=\frac{1}{4}$) for $g=2$ and to full sibs for $g=1$ (provided we set $\phi_n(l)=0$, $n\leqslant 0$).

(c) *Cousins*

In Fig. 1, **E** and **F** are full sibs, and so their respective offspring **H** and **I** are (first or full) cousins. They may share one or zero pairs of alleles ibd with probabilities $k_1=\frac{1}{4}$ and $k_0=\frac{3}{4}$. The haplotypes that they receive from their sibling parents may each be non-recombinant, with probability $(1-c)^2$, in which case they carry ibd alleles at each locus with probability $[\frac{1}{4}(1-c)^2+\frac{1}{4}c^2]$. Alternatively, the haplotypes that they receive from their sibling parents may each be recombinant, with probability c^2, in which case they carry ibd alleles at each locus with probability $\frac{1}{8}$. Therefore,

$$\begin{aligned}\Pr(\breve{k}_{1i}\breve{k}_{1j})&=\tfrac{1}{2}(1-c)^2[\tfrac{1}{2}(1-c)^2+\tfrac{1}{2}c^2]+\tfrac{1}{8}c^2\\&=8b^4-4b^3+\tfrac{3}{2}b^2-\tfrac{1}{2}b+\tfrac{1}{8}\end{aligned}\qquad(13)$$

and hence

$$\begin{aligned}&\mathrm{Var}_{\mathrm{FC}}(\breve{k}_1,l)=8\phi_4(l)-4\phi_3(l)+\tfrac{3}{2}\phi_2(l)-\tfrac{1}{2}\phi_1(l),\\&\mathrm{Var}_{\mathrm{FC}}(\breve{R},l)=2\phi_4(l)-\phi_3(l)+\tfrac{3}{8}\phi_2(l)-\tfrac{1}{8}\phi_1(l).\end{aligned}\qquad(14)$$

Note that the variances differ from those for great uncle–great nephew, although they have the same relationship parameters k_1 and R.

(d) *Descendants of cousins*

In Fig. 1, **H** and **L** are cousins once removed. An individual shares a haplotype with the offspring of a cousin only if the cousin transmits it without recombination. Hence, the joint probability of sharing is b times that for cousins. Setting $g=3$ for cousins ($R=\frac{1}{8}$), so $g=4$ for cousins once removed, $g=5$ for second cousins and for cousins twice removed and $g=6$ for third cousins, The variances are

$$\begin{aligned}\mathrm{Var}_{C,g}(\breve{k}_1, l) &= 8\phi_{g+1}(l) - 4\phi_g(l) + \tfrac{3}{2}\phi_{g-1}(l) \\ &\quad - \tfrac{1}{2}\phi_{g-2}(l) + \tfrac{1}{8}\phi_{g-3}(l), \\ \mathrm{Var}_{C,g}(\breve{R}, l) &= 2\phi_{g+1}(l) - \phi_g(l) + \tfrac{3}{8}\phi_{g-1}(l) \\ &\quad - \tfrac{1}{8}\phi_{g-2}(l) + \tfrac{1}{32}\phi_{g-3}(l),\end{aligned}$$

and also $\mathrm{Var}_{C,1}(\breve{R}, l) = \mathrm{Var}_{FS}(\breve{R}, l)$.

(v) *Bilineal relatives*

(a) *General methodology*

Bilineal relatives can receive identical alleles from each of the two different pedigrees. Full sibs have two parents in common and each may transmit identical alleles to the sibs. Double first cousins have two pairs of grandparents in common, and each pair may transmit identical alleles to the cousins. It is convenient to refer to the two pedigrees as 'maternal' and 'paternal', although this may not be the case for double first cousins. In Fig. 1, **E** and **F** are full sibs and can receive identical alleles from each of their parents **B** and **C**. If **M** and **N** are also full sibs, then **H** and **I** are double first cousins who may receive ibd alleles from both sets of grandparents, namely **B**, **C** and the parents of **M**, **N**.

Using superscripts m, p for maternal and paternal events in order to extend the previous definitions of actual identity indicators, the required indicators can be partitioned as

$$\begin{aligned}\breve{k}_2 &= \breve{k}_1^{\,m}\breve{k}_1^{\,p}, \\ \breve{k}_1 &= \breve{k}_1^{\,m}(1-\breve{k}_1^{\,p}) + (1-\breve{k}_1^{\,m})\breve{k}_1^{\,p}, \\ \breve{k}_0 &= (1-\breve{k}_1^{\,m})(1-\breve{k}_1^{\,p}).\end{aligned}$$

As we assume no inbreeding, $\breve{k}_1^{\,m}$ and $\breve{k}_1^{\,p}$ are independent and have expected values denoted $\alpha^m = k_1^{\,m}$ and $\alpha^p = k_1^{\,p}$. Therefore, $k_2 = \alpha^m\alpha^p$, $k_1 = \alpha^m(1-\alpha^p) + (1-\alpha^m)\alpha^p$ and $k_0 = (1-\alpha^m)(1-\alpha^p)$. For full sibs, for example, $\alpha^m = \alpha^p = \frac{1}{2}$, $k_2 = k_0 = \frac{1}{4}$ and $k_1 = \frac{1}{2}$.

Hence, the variance of the actual relationship, $\breve{R} = \frac{1}{2}\left(\breve{k}_1^{\,m} + \breve{k}_1^{\,p}\right)$, can be written in an alternative form to eqn (1) as

$$\mathrm{Var}(\breve{R}) = \tfrac{1}{4}[\alpha^m(1-\alpha^m) + \alpha^p(1-\alpha^p)]. \qquad (15)$$

The sharing of either or both maternal and paternal alleles can extend to each of the two loci, i and j, and we introduce the expected products

$$\beta^m = E(\breve{k}_{1i}^{\,m}\breve{k}_{1j}^{\,m}), \quad \beta^p = E(\breve{k}_{1i}^{\,p}\breve{k}_{1j}^{\,p}).$$

For full sibs, these values are each the same as for sharing of alleles transmitted from their common parent to half-sibs (eqn (6)), $\beta^m = \beta^p = \frac{1}{2}[(1-c)^2 + c^2]$.

As maternal and paternal alleles are inherited independently,

$$E(\breve{k}_{1i}^{\,m}\breve{k}_{1i}^{\,p}) = E(\breve{k}_{1i}^{\,m}\breve{k}_{1j}^{\,p}) = E(\breve{k}_{1j}^{\,m}\breve{k}_{1i}^{\,p}) = (\breve{k}_{1j}^{\,m}\breve{k}_{1j}^{\,p}) = \alpha^m\alpha^p.$$

The expected product of sharing two pairs of alleles at two loci for bilineal relatives is

$$\begin{aligned}E(\breve{k}_{2i}\breve{k}_{2j}) &= E(\breve{k}_{1i}^{\,m}\breve{k}_{1i}^{\,p}\breve{k}_{1j}^{\,m}\breve{k}_{1j}^{\,p}) \\ &= E(\breve{k}_{1i}^{\,m}\breve{k}_{1j}^{\,m})\ E(\breve{k}_{1i}^{\,p}\breve{k}_{1j}^{\,p}) = \beta^m\beta^p\end{aligned}$$

and the covariance of the double-sharing indicators is

$$\mathrm{Cov}(\breve{k}_{2i}, \breve{k}_{2j}) = E(\breve{k}_{2i}\hat{k}_{2j}) - E(\breve{k}_{2i})E(\breve{k}_{2j}) = \beta^m\beta^p - (\alpha^m\alpha^p)^2.$$

For the other covariances, we note that terms such as $[E(\breve{k}_i^{\,m}\breve{k}_j^{\,p}) - E(\breve{k}_i^{\,m})E(\breve{k}_j^{\,p})]$ contribute zero, whereas terms such as $[E(\breve{k}_i^{\,m}\breve{k}_j^{\,m}) - E(\breve{k}_i^{\,m})E(\breve{k}_j^{\,m})]$ contribute $(\beta_{ij}^{\,m} - \alpha_i^{\,m}\alpha_j^{\,m})$. The remaining covariances are obtained similarly. The covariance $\mathrm{Cov}(\breve{k}_{1i}, \breve{k}_{1j})$ comprises four terms: from sharing of both paternal alleles but neither maternal allele and *vice versa*, and from sharing of paternal but not maternal alleles at the first locus and of maternal but not paternal alleles at the second locus and *vice versa*. It is convenient to define $\omega^m = \beta^m - (\alpha^m)^2$ and $\omega^p = \beta^p - (\alpha^p)^2$.

We obtain

$$\mathrm{Cov}(\breve{k}_{2i}, \breve{k}_{2j}) = (\alpha^m)^2\omega^p + (\alpha^p)^2\omega^m + \omega^m\omega^p$$

and also

$$\begin{aligned}&\mathrm{Cov}(\breve{k}_{1i}, \breve{k}_{1j}) = (1-2\alpha^m)^2\omega^p + (1-2\alpha^p)^2\omega^m + 4\omega^m\omega^p, \\ &\mathrm{Cov}(\breve{k}_{0i}, \breve{k}_{0j}) = (1-\alpha^m)^2\omega^p + (1-\alpha^p)^2\omega^m + \omega^m\omega^p, \\ &\mathrm{Cov}(\breve{k}_{2i}, \breve{k}_{1j}) + \mathrm{Cov}(\breve{k}_{1i}, \breve{k}_{2j}) = 2\alpha^m(1-2\alpha^m)\omega^p \\ &\quad + 2\alpha^p(1-2\alpha^p)\omega^m - 4\omega^m\omega^p, \\ &\mathrm{Cov}(\breve{k}_{2i}, \breve{k}_{0j}) + \mathrm{Cov}(\breve{k}_{0i}, \breve{k}_{2j}) = -2\alpha^m(1-\alpha^m)\omega^p \\ &\quad - 2\alpha^p(1-\alpha^p)\omega^m + 2\omega^m\omega^p, \\ &\mathrm{Cov}(\breve{k}_{1i}, \breve{k}_{0j}) + \mathrm{Cov}(\breve{k}_{0i}, \breve{k}_{1j}) = -2(1-\alpha^m)(1-2\alpha^m)\omega^p \\ &\quad - 2(1-\alpha^p)(1-2\alpha^p)\omega^m - 4\omega^m\omega^p.\end{aligned} \qquad (16)$$

Note that these six expressions sum to zero, as $\breve{k}_0 + \breve{k}_1 + \breve{k}_2 = 1$ at each locus. For unlinked loci, $\beta^m = (\alpha^m)^2$ and $\beta^p = (\alpha^p)^2$, all these expressions (16) are zero. For completely linked loci, $\beta^m = \alpha^m$ and $\beta^p = \alpha^p$, the covariances reduce to the variances and covariances of the single-locus indicators.

Averaging over just two loci, i, j:

$$\breve{R}=\tfrac{1}{2}(\breve{k}_{2i}+\breve{k}_{2j})+\tfrac{1}{4}(\breve{k}_{1i}+\breve{k}_{1j}).$$

Using one-locus results and the two-locus covariances in this case

$$\mathrm{Var}(\breve{R})=\tfrac{1}{8}[\omega^{\mathrm{m}}+\omega^{\mathrm{p}}+\alpha^{\mathrm{m}}(1-\alpha^{\mathrm{m}})+\alpha^{\mathrm{p}}(1-\alpha^{\mathrm{p}})].$$

As expected, this does not involve the product $\omega^{\mathrm{m}}\omega^{\mathrm{p}}$ (or, equivalently, $\beta^{\mathrm{m}}\,\beta^{\mathrm{p}}$) because the maternal and paternal alleles are transmitted independently. For unlinked loci, the variance is half the single-locus value shown in eqn (15).

(b) *Full sibs*

For full sibs, $\alpha^{\mathrm{m}}=\alpha^{\mathrm{p}}=\alpha=\tfrac{1}{2}$, $\beta^{\mathrm{m}}=\beta^{\mathrm{p}}=\tfrac{1}{2}[(1-c)^2+c^2]=4b^2-2b+\tfrac{1}{2}$. Therefore $\beta^{\mathrm{m}}\beta^{\mathrm{p}}=16b^4-16b^3+8b^2-2b+\tfrac{1}{4}$, which equals 1/16 when $c=\tfrac{1}{2}$. Using $\mathrm{Cov}(\breve{k}_{2i},\breve{k}_{2j})=\beta^{\mathrm{m}}\beta^{\mathrm{p}}-(\alpha^{\mathrm{m}}\alpha^{\mathrm{p}})^2$ from eqns (16) and integrating over a chromosome of length l:

$$\begin{aligned}
\mathrm{Var}_{\mathrm{FS}}(\breve{k}_2,l)&=\mathrm{Var}_{\mathrm{FS}}(\breve{k}_0,l)=16\phi_4(l)-16\phi_3(l)\\
&\quad+8\phi_2(l)-2\phi_1(l),\\
\mathrm{Var}_{\mathrm{FS}}(\breve{k}_1,l)&=64\phi_4(l)-64\phi_3(l)+24\phi_2(l)-4\phi_1(l),\\
\mathrm{Cov}_{\mathrm{FS}}(\breve{k}_2,\breve{k}_1,l)&=\mathrm{Cov}_{\mathrm{FS}}(\breve{k}_1,\breve{k}_0,l)=-32\phi_4(l)\\
&\quad+32\phi_3(l)-12\phi_2(l)+2\phi_1(l),\\
\mathrm{Cov}_{\mathrm{FS}}(\breve{k}_2,\breve{k}_0,l)&=16\phi_4(l)-16\phi_3(l)+4\phi_2(l).
\end{aligned}$$

An alternative summary of these expressions is given in Box 1. The variance of the actual relationship for full sibs can be obtained from these results, and is $\mathrm{Var}_{\mathrm{FS}}(\breve{R},l)=2\phi_2(l)-\phi_1(l)$, i.e. twice that for half-sibs, as noted previously. The variance of $\breve{k}_2$ was derived by Visscher *et al.* (2006), who also pointed out that $\mathrm{Cov}_{\mathrm{FS}}(\breve{R},\breve{k}_2,l)=\mathrm{Var}_{\mathrm{FS}}(\breve{R},l)$. The regression of $\breve{k}_2$ on $\breve{R}$ is therefore 1·0. The genetic covariance of phenotypes of quantitative traits of relatives (ignoring epistasis) is given by $RV_{\mathrm{A}}+k_2V_{\mathrm{D}}$, where V_{A} and V_{D} are the additive and dominance variances (Falconer & Mackay, 1996) and traditionally pedigree relationships are used. Estimates of the additive genetic and dominance variances free of environmental covariances for quantitative traits can be obtained by regressing the resemblance of trait values of full sibs to their actual genome shared, $\breve{R}V_{\mathrm{A}}+\breve{k}_2V_{\mathrm{D}}$, if dense markers are available. The estimates of V_{A} and V_{D} are therefore highly correlated, however (Visscher *et al.*, 2006).

(c) *Double first cousins*

For double first cousins $\alpha^{\mathrm{m}}=\alpha^{\mathrm{p}}=\tfrac{1}{4}$ and, utilizing the results for descendants of first cousins (eqn (13)), $E(\breve{k}_{1i}\breve{k}_{1j})=8b^4-4b^3+\tfrac{3}{2}b^2-\tfrac{1}{2}b+\tfrac{1}{8}$, it follows that

$$\begin{aligned}
\mathrm{Var}_{\mathrm{DFC}}(\breve{k}_2,l)&=64\phi_8(l)-64\phi_7(l)+40\phi_6(l)-20\phi_5(l)\\
&\quad+\tfrac{33}{4}\phi_4(l)-\tfrac{5}{2}\phi_3(l)+\tfrac{5}{8}\phi_2(l)-\tfrac{1}{8}\phi_1(l).
\end{aligned}$$

The other variances and covariances can be expressed simply (Box 1) in terms of $\mathrm{Var}_{\mathrm{DFC}}(\breve{k}_2,l)$ and $\mathrm{Var}_{\mathrm{FC}}(\breve{k}_1,l)=8\phi_4(l)-4\phi_3(l)+\tfrac{3}{2}\phi_2(l)-\tfrac{1}{2}\phi_1(l)$.

The variance of the actual relationship is double that of first cousins:

$$\begin{aligned}
\mathrm{Var}_{\mathrm{DFC}}(\breve{R},l)&=2\mathrm{Var}_{\mathrm{FC}}(\breve{R},l)=4\phi_4(l)-2\phi_3(l)\\
&\quad+\tfrac{3}{4}\phi_2(l)-\tfrac{1}{4}\phi_1(l).
\end{aligned}$$

Also $\mathrm{Cov}_{\mathrm{DFC}}(\breve{R},\breve{k}_2,l)=\tfrac{1}{2}\mathrm{Var}_{\mathrm{FC}}(\breve{R},l)$, and so the regression of $\breve{k}_2$ on $\breve{R}_2$ is one-half.

(d) *Mothers full sibs, fathers first cousins*

The method that we have established allows for asymmetry in the two pedigrees that lead to sets of identical alleles for a pair of relatives. If, for example, the mothers are full sibs and the fathers are first cousins $\alpha^{\mathrm{m}}=\tfrac{1}{2}$, $\alpha^{\mathrm{p}}=\tfrac{1}{4}$, $\beta^{\mathrm{m}}=4b^2-2b+\tfrac{1}{2}$ and $\beta^{\mathrm{p}}=8b^4-4b^3+\tfrac{3}{2}b^2-\tfrac{1}{2}b+\tfrac{1}{8}$. The results then follow.

(vi) *Sex-related phenomena*

(a) *Differences in map length between sexes*

In the analysis we have assumed that the map distance is the same in both sexes. Typically, however, the sexes differ in map length, i.e. in the rate of recombination per unit of physical length of the genome. For humans, the autosomal map length in females is 44 M approximately and in males 28 M (Kong *et al.*, 2004), with the male/female ratio ranging among autosomes from 57 to 85%, typically differing more for the longer chromosomes. We quantify the impact on the variation in genome sharing on the sex through which transmission occurs.

It would be possible to restructure the analysis and specify a ratio of map to physical length for each chromosome and integrate an extension to eqn (4) over physical rather than map length. For maintaining the same notation as previously, however, we simply assume that the sex-averaged map length for a particular chromosome is l, but the map length in females is given by $l(1+\lambda)$ and in males by $l(1-\lambda)$. Initially we take a more general approach, and assume that the map length for transmissions at generation i is given by la_i and that recombination fractions between any pair of sites are functions of la_i. Thus, for a pair of loci d M apart on the sex-averaged linkage map and assuming Haldane's mapping function, their recombination fraction is $\tfrac{1}{2}(1-\mathrm{e}^{-2da_i})$, $0<d<l$, at generation i. We consider lineal relationships.

Equations (4*a*) and (4*b*) for $\phi_n(l)$ can now be generalized:

$$\phi_n^{\circ}(l;a_1,\ \ldots,a_n)=\frac{2}{l^2}\left(\frac{1}{4}\right)^n\int_{x=0}^{l}\int_{y=0}^{x}\left[\prod_{i=1}^{n}(1+\mathrm{e}^{-2(x-y)a_i})-1\right]\mathrm{d}y\ \mathrm{d}x$$
$$=\frac{2}{l^2}\left(\frac{1}{4}\right)^n\int_{x=0}^{l}\int_{y=0}^{x}\left[\sum_{\delta_1=0}^{1}\sum_{\delta_2=0}^{1}\cdots\sum_{\delta_n=0}^{1}\mathrm{e}^{-2(x-y)\sum_{i=1}^{n}a_i\delta_i}\right]\mathrm{d}y\ \mathrm{d}x,\quad \sum_{i=1}^{n}\delta_i\neq 0, \tag{17a}$$

$$=\frac{1}{2l^2}\left(\frac{1}{4}\right)^n\left[\sum_{\delta_1=0}^{1}\sum_{\delta_2=0}^{1}\cdots\sum_{\delta_n=0}^{1}\left(\frac{2l}{\sum_{i=1}^{n}a_i\delta_i}-\frac{1}{(\sum_{i=1}^{n}a_i\delta_i)^2}+\frac{\mathrm{e}^{-2l\sum_{i=1}^{n}a_i\delta_i}}{(\sum_{i=1}^{n}a_i\delta_i)^2}\right)\right],\quad \sum_{i=1}^{n}\delta_i\neq 0, \tag{17b}$$

If $a_i=1$ for all i, eqns (17*a*) reduce to (4*a*) and (17*b*) to (4*b*).

Although (17*b*) can be used directly, we now simplify for the case where there are just two values of a_i, namely $1\pm\lambda$. Assume that m of the n transmissions are through males, with $n-m$ correspondingly through females, and extend the definition of $\phi_n(l)$ accordingly as $\phi^*_{n,m}(l,\lambda)$. The sequence in which male or female transmissions occur does not matter. The expansion of the summations in (17*b*) involves terms with $r=\sum_{i=1}^{n}\delta_i$ terms in the sum $\sum_{i=1}^{n}a_i\delta_i$ and of these r there are, say, s transmissions through males, where $\max(0,r-n+m)\leqslant s\leqslant\min(m,r)$. Hence $\sum_{i=1}^{n}a_i\delta_i=r+(r-2s)\lambda=\rho$, say, and

$$\phi^*_{n,m}(l,\lambda)=\frac{1}{2l^2}\left(\frac{1}{4}\right)^n\sum_{r=1}^{n}\sum_{s=\max(0,r-n+m)}^{\min(m,r)}\binom{m}{s}\binom{n-m}{r-s}\times\left[2\rho-\frac{1}{\rho^2}+\frac{\mathrm{e}^{-2l/\rho}}{\rho^2}\right], \tag{18}$$

i.e. ρ replaces r in (4) and hypergeometric coefficients in s are introduced. The n generations here do not include that of the initial transmission from parent to offspring, but those starting with the subsequent transmission to grandoffspring, so

$$\mathrm{Var}_{\mathrm{Lin},g,m}(\breve{k}_1,l,\lambda)=\phi^*_{g-1,m}(l,\lambda).$$

For collateral relatives and their offspring, the general formulation can be extended. For example, for a pair of paternal half-sibs

$$\mathrm{Var}_{\mathrm{HS}}(\breve{R},l,\lambda)=\phi_2(l(1-\lambda))-\tfrac{1}{2}\phi_1(l(1-\lambda))$$
$$=\phi^*_{2,2}(l,\lambda)-\tfrac{1}{2}\phi^*_{1,1}(l,\lambda)$$

and for a pair of half-cousins, whose mothers were paternal half-sibs,

$$\mathrm{Var}_{\mathrm{HC}}(\breve{R},l,\lambda)=\phi^*_{4,2}(l,\lambda)-\tfrac{1}{2}\phi^*_{3,1}(l,\lambda)+\tfrac{1}{8}\phi^*_{2,0}(l,\lambda).$$

As both sexes of parents contribute to resemblance among full sibs, the differences in map length have much impact only in later generations.

For humans, λ averages approximately 0·25, and we illustrate the calculations for a chromosome with $l=1$ M. For $n=2$, i.e. great grandparent–great grandoffspring (with the sex of the great grandparent irrelevant), $\phi^*_{2,0}(1,0{\cdot}25)=\phi_2(1{\cdot}25)=0{\cdot}0833$, $\phi^*_{2,1}(1,0{\cdot}25)=0{\cdot}0954$ and $\phi^*_{2,2}(1,0{\cdot}25)=\phi_2(0{\cdot}75)=0{\cdot}1088$. The corresponding standard deviations (SDs) of k_1 are 0·289, 0·309 and 0·330, describing subsequent transmissions twice through females, once through each sex, and twice through males, respectively. For $n=4$, $m=0,1,\ldots,4$, $\phi^*_{n,m}(1,0{\cdot}25)=0{\cdot}0197$, 0·0214, 0·0231, 0·0251 and 0·0272, respectively.

It is straightforward to evaluate eqn (18) directly. These examples illustrate, however, that linear interpolations can provide good approximations. One alternative is to interpolate on ϕ using $(1-m/n)\phi_n\times(l(1+\lambda))+(m/n)\phi_n(l(1-\lambda))$, which for the example above for $n=4$ and $m=1$, 2 and 3 gives 0·0214, 0·0235 and 0·0253, respectively. Another is to interpolate on l using $\phi_n(l(1+(1-2m/n)\lambda))$, for which corresponding values are 0·0212, 0·0229 and 0·0249.

(b) *Sex limited recombination*

For species such as *Drosophila melanogaster* there is no recombination in males, so autosomes are transmitted intact to the offspring and the variance in sharing with and among their descendants is increased. The probability that a parental pair of genes is transmitted to an offspring is $\frac{1}{2}(1-c)$ through a female and $\frac{1}{2}$ through a male. If m of the $n=g-1$ transmissions to descendants after the first generation (as the sex of the ancestor is not relevant) are through a male,

$$\mathrm{cov}(\breve{k}_{1i},\breve{k}_{1j})=(\tfrac{1}{2})^m[\tfrac{1}{2}(1-c)]^{n-m}-(\tfrac{1}{4})^n$$
$$=(\tfrac{1}{2})^m\{[\tfrac{1}{2}(1-c)]^{n-m}-(\tfrac{1}{4})^{n-m}\}+(\tfrac{1}{4})^{n-m}[(\tfrac{1}{2})^m-(\tfrac{1}{4})^m].$$

Hence,

$$\mathrm{Var}_{\mathrm{Lin(SL)}\,g,m}(\breve{k}_1,l)=(\tfrac{1}{2})^m\{\phi_{g-1-m}(l)+(\tfrac{1}{4})^{g-1-m}[1-(\tfrac{1}{2})^m]\}.$$

To take another example: for full sibs, the probability of sharing is $\frac{1}{2}$ for genes from their father and $\frac{1}{2}[(1-c)^2+c^2]$ from their mother. Therefore, by summing components for maternal and paternal half-sibs, $\mathrm{Var}_{\mathrm{FS(SL)}}(\breve{R},l)=\phi_2(l)-\frac{1}{2}\phi_1(l)+\frac{1}{16}$.

(c) *Sex chromosomes*

Previous formulae apply for the autosomes and we now consider the sex chromosomes (assuming mammalian X, Y sex determination and ignoring the pseudo-autosomal region). For the Y chromosome, father and son share a genome exactly and there is no variation in sharing. Father and son do not share an X chromosome, and so for lineal descendants any male–male transmission in the pathway results in no sharing of descendant with the ancestor. A daughter receives a copy of her father's X chromosome without sampling, and so any male to female transmission reduces by one the number of generations of sampling in eqn (2). Son and daughter receive an X from their mother with recombination as for the autosomes. We consider only the case of full sibs in detail, but sampling variances for genome sharing on the X chromosome can be deduced for any relationship. Visscher (2009) gives further discussion for sex-linked chromosomes.

We retain the k_1^{m}, k_1^{p} notation for the ibd of maternal or paternal alleles, adding a subscript to indicate X-linkage. For two full brothers, $k_{1\mathrm{X}}^{\mathrm{p}}$ is not defined and $k_{1\mathrm{X}}^{\mathrm{m}}=\frac{1}{2}$, the same as k_1 for half-sibs; $\breve{k}_{2\mathrm{X}}=0$, $\breve{k}_{1\mathrm{X}}=\breve{k}_{1\mathrm{X}}^{\mathrm{m}}$ and $\breve{k}_{0\mathrm{X}}=1-\breve{k}_{1\mathrm{X}}^{\mathrm{m}}$. Integrating over the X chromosome of length l_{X} gives $\mathrm{Var}_{\mathrm{BB}}(\breve{k}_{1\mathrm{X}})=4\phi_2(l_{\mathrm{X}})-2\phi_1(l_{\mathrm{X}})$, using the autosomal result for half-sibs (11). For a sister and brother, $k_{1\mathrm{X}}^{\mathrm{p}}$ is still not defined and $k_{1\mathrm{X}}^{\mathrm{m}}$ is as for half-sibs with a value of $\frac{1}{2}$. Hence, $\mathrm{Var}_{\mathrm{BS}}(\breve{k}_{1\mathrm{X}})=\mathrm{Var}_{\mathrm{BB}}(\breve{k}_{1\mathrm{X}})$. For two sisters, $\breve{k}_{1\mathrm{X}}^{\mathrm{p}}=1$ and $\breve{k}_{1\mathrm{X}}^{\mathrm{m}}$ is as for half-sibs. From the previous results, $\breve{k}_{2\mathrm{X}}=\breve{k}_{1\mathrm{X}}^{\mathrm{m}}$, $\breve{k}_{1\mathrm{X}}=1-\breve{k}_{1\mathrm{X}}^{\mathrm{m}}$ and $\breve{k}_{0\mathrm{X}}=0$; therefore, $\mathrm{Var}_{\mathrm{SS}}(\breve{k}_{2\mathrm{X}})=\mathrm{Var}_{\mathrm{SS}}(\breve{k}_{1\mathrm{X}})=-\mathrm{Cov}_{\mathrm{SS}}(\breve{k}_{2\mathrm{X}},\breve{k}_{1\mathrm{X}})=4\phi_2(l_{\mathrm{X}})-2\phi_1(l_{\mathrm{X}})$.

(vii) *Examples*

Examples of the SDs of actual proportion of genome shared ($\breve{k}_2+\frac{1}{2}\breve{k}_1$) as a function of map length for single chromosomes are given in Fig. 2*a* for descendants of full sibs. It is noticeable that there remains a substantial variation even for the longest chromosomes illustrated (4 Morgans), i.e. longer than most chromosomes in most species. Although the SD becomes smaller as the individuals become less related, the CV becomes larger (Fig. 2*b*) (Visscher, 2009). Indeed the CV exceeds unity for all but close relationships, even for chromosomes of map length 2 M.

Comparisons between lineal descendants and those of half- and full sibs are given in Fig. 3 for two examples of relationship. With complete linkage the variance depends only on relationship (Table 1). Although the differences are quite small, with increasing map length the variance declines less rapidly with increased chromosome length for lineal descendants than for those involving half sibs, which in turn

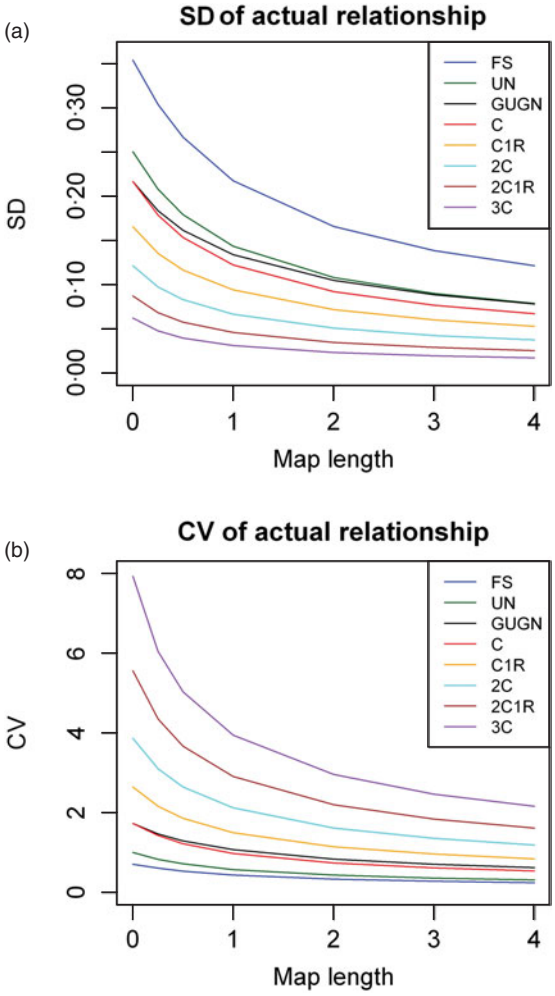

Fig. 2. (*a*) SD and (*b*) CV of actual relationship (proportion of genome shared, $\breve{R}=\breve{k}_2+\frac{1}{2}\breve{k}_1$), for a single chromosome as a function of map length and relationship for full sibs (FS) and their descendants: uncle nephew (UN), cousins (C), cousins once removed (C1R), second cousins (2C), second cousins once removed (2C1R) and third cousins (3C).

show a faster decline than descendants of full sibs (Fig. 3). This is presumably because the latter can be ibd at a pair of loci on a pair of recombinant chromosomes: terms in c^2 appearing in eqns (6) and (13), for example, but not in (2). Great uncle–nephew and first cousins, which have the same relationship, differ in the variance of sharing, but not very much (Fig. 3).

For a mammalian or avian genome with multiple chromosomes, the variation and skew are reduced. Taking data for human autosomes from Kong *et al.* (2004), we assumed that the 22 chromosomes could be grouped into six classes each of 2–8 chromosomes, each member of which was of similar map and genome length, as follows: (1–2) 2·75 M, (3–6) 2·10 M, (7–12) 1·75 M, (13–20) 1·25 M, (21–22) 0·75 M. Results are given in Table 2 for a wide range of relationships. The results are, however, little different from what would be expected from the same number

Table 2. *SD of actual relationship* ($\breve{R}=\breve{k}_2+\frac{1}{2}\breve{k}_1$) *for a model human genome for different pedigree relationships* ($R=2\theta$)

	Lineal descendants		Half sibs' descendants		Full sibs' descendants		
R	Relationship[a]	SD($\breve{R}$)[b]	Relationship	SD($\breve{R}$)[b]	Relationship	SD($\breve{R}$)[b]	SD($\breve{R}$)[c]
0·5	P–O	0·0			FS	0·0392	0·0384
0·25	GP–GO	0·0362	HS	0·0277	UN	0·0256	0·0251
0·125					GUGN	0·0247	0·0241
0·125	GGP–GGO	0·0291	HUN	0·0256	C	0·0218	0·0214
0·0625	G3P–G3O	0·0206	HC	0·0188	C1R	0·0170	0·0166
0·0312	G4P–G4O	0·0139	HC1R	0·0130	2C	0·0120	0·0117
0·0156	G5P–G5O	0·0093	HSC	0·0087	2C1R	0·0082	0·0080
0·0078	G6P–G6O	0·0062	HSC1R	0·0058	3C	0·0055	0·0054

[a] P–O, parent–offspring; G*n*P–G*n*O, great(*n*)grandparent – great(*n*)grandoffspring; H, half; UN, uncle–nephew; GUGN, great uncle–great nephew; C, 2C, 3C first, second, third cousin; 1R, once removed.
[bc] SD($\breve{R}$) computed assuming 22 chromosomes: [b]with differing map lengths, total 35·9 M (see text), [c]each of length $\Sigma 35{\cdot}9/22=1{\cdot}63$ M.

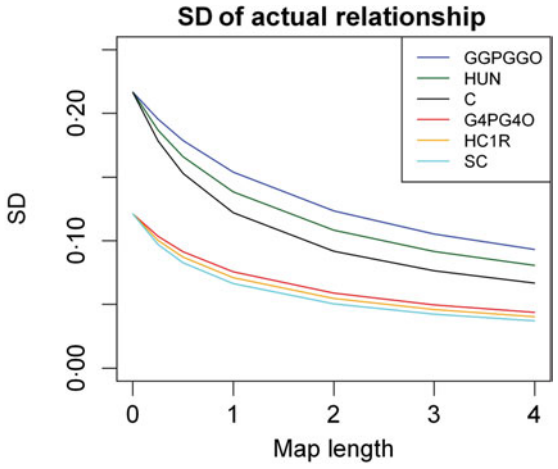

Fig. 3. SD of actual relationship (proportion of genome shared, $\breve{R}=\breve{k}_2+\frac{1}{2}\breve{k}_1$), for a single chromosome as a function of map length and relationship for three different pedigrees for two different pedigree relationships: $R=0{\cdot}125$: great grandparent–great grandoffspring (GGPGGO), half-uncle–nephew (HUN), great uncle–great nephew (GUGN), cousins (C); and $R=0{\cdot}03125$: greatgreatgreat grandparent–GGGGoffspring (G4PG4O), half–cousins once removed (HC1R) and second cousins (2C).

of chromosomes each of the average map length, as shown by an example in the last column of Table 2 and as pointed out previously (Hill, 1993*a*; Visscher, 2009). The average chromosomal length is about 1·6 Morgans, so with 22 chromosomes, the SD, CV and skew of sharing are approximately 20% of those for individual chromosomes.

3. Skew of the distribution of genome sharing

(i) *Methods*

The methods that we have used for evaluating the variance of actual identity can be extended for dealing with higher moments, although the algebra becomes increasingly prohibitive. Here, we consider the magnitude of skew, initially giving formulae for individual genes.

The third central moment of an allele sharing indicator variable $\breve{k}_m$, $m=0$, 1, 2, is

$$E[(\breve{k}_m-k_m)^3]=E(\breve{k}_m^3)-3k_m\mathrm{Var}(\breve{k}_m)-k_m^3$$
$$=k_m(1-k_m)(1-2k_m)$$

and the corresponding skew coefficient is

$$\gamma_1(\breve{k}_m)=\frac{\mu_3(\breve{k}_m)}{[\mu_2(\breve{k}_m)]^{3/2}}=\frac{(1-2k_m)}{\sqrt{k_m(1-k_m)}}.$$

The $\breve{k}_m$s are symmetrically distributed if they are equal 0·5 and positively skewed if less than 0·5. The third central moment of the actual relationship can be shown to be

$$\mu_3(\breve{R})=E[(\breve{R}-R)^3]=(1-2R)[R(1-R)-\tfrac{3}{8}k_1]$$

For lineal descendants, i.e. $k_2=0$, $\gamma_1(\breve{R})=\gamma_1(\breve{k})$ and the distribution of actual relationship or co-ancestry is symmetric if $k_1=0{\cdot}5$, e.g. grandparent–grand offspring, half–sibs and uncle–nephew. The distribution of R is also symmetric for full sibs.

For evaluating the skew in genome sharing, we extend the methods used in order to compute the variance in actual relationship, but in view of the complexity of the analysis, restrict it to the case of lineal descendants (i.e. $k_2=0$ at all loci). Thus, we evaluate $E(\breve{k}_1^3)$ as an average over r loci, where r becomes infinitely large:

$$E(\breve{k}_1^3)=\frac{1}{r^3}E\left(\sum_h\sum_i\sum_j\breve{k}_{1h}\breve{k}_{1i}\breve{k}_{1j}\right)$$

Consider the expected value of allele sharing $E(\breve{k}_{1h}\breve{k}_{1i}\breve{k}_{1j})$ at three loci h, i, j so ordered along a chromosome. A three-locus haplotype is transmitted intact from parent to offspring with probability $\frac{1}{2}(1-c_1)(1-c_2)$, where c_1 and c_2 are the recombination fractions between loci h, i and i, j, respectively. The probability is $\frac{1}{8}$ if the loci are unlinked. The probability of allele sharing for three-linked loci between two individuals, one of which is a g-generation lineal descendent of the other, is therefore

$$E(\breve{k}_{1h}\breve{k}_{1i}\breve{k}_{1j}) = \left(\tfrac{1}{2}\right)^{g-1}(1-c_1)^{g-1}(1-c_2)^{g-1}. \tag{19}$$

This equation extends the two-locus result in eqn (2) and can be evaluated over each chromosome by invoking Haldane's mapping function to write recombination fractions in terms of map lengths and integrating:

$$E(\breve{k}_{1h}\breve{k}_{1i}\breve{k}_{1j}) - k_1^3 = \frac{6}{l^3}\left(\frac{1}{2}\right)^{3(g-1)}\int_0^l\int_0^x\int_0^y [(1+e^{-2(x-y)})^{g-1} \times (1+e^{-2(y-z)})^{g-1} - 1]\,dz\,dy\,dx. \tag{20}$$

As the analysis has also to deal with descendants of collateral relatives, we generalize the integration, illustrating the process for half-sibs. The probability that a pair of half-sibs share an allele ibd at each of the three loci is

$$E(\breve{k}_{1h}\breve{k}_{1i}\breve{k}_{1j}) = \tfrac{1}{2}\{[1-c_1)(1-c_2)]^2 + [(1-c_1)c_2]^2 + [c_1(1-c_2)]^2 + [c_1c_2]^2\}.$$

In order to evaluate this expression, we expand it in terms of $(1-c_1)$ and $(1-c_2)$:

$$\begin{aligned} E(\breve{k}_{1h}\breve{k}_{1i}\breve{k}_{1j}) = \tfrac{1}{2}[&4(1-c_1)^2(1-c_2)^2 - 4(1-c_1)^2(1-c_2)^1 \\ &- 4(1-c_1)^1(1-c_2)^2 + 4(1-c_1)^1(1-c_2)^1 \\ &+ 2(1-c_1)^2(1-c_2)^0 + 2(1-c_1)^0(1-c_2)^2 \\ &- 2(1-c_1)^1(1-c_2)^0 - 2(1-c_1)^0(1-c_2)^1 + 1]. \end{aligned} \tag{21}$$

As some terms have different exponents for $(1-c_1)$ and $(1-c_2)$, we redefine the integral more generally than shown in eqn (20), and the exponents are not generation numbers *per se*. We express $(1-c_1)^m(1-c_2)^n$ in terms of map distances and define

$$\Phi_{m,n}(l) = \frac{6}{l^3}\left(\frac{1}{2}\right)^{m+n}\int_0^l\int_0^x\int_0^y [(1+e^{-2(x-y)})^m \times (1+e^{-2(y-z)})^n - 1]\,dz\,dy\,dx$$

$$\begin{aligned} = \left(\frac{1}{2}\right)^{m+n}\Bigg\{&1 + \sum_{i=1}^{m}\binom{m}{i}\frac{2i^2l^2 - 2il + 1 - e^{-2il}}{8i^3} \\ &+ \sum_{j=1}^{n}\binom{n}{j}\frac{2j^2l^2 - 2jl + 1 - e^{-2jl}}{8j^3} \\ &+ \sum_{i=1}^{\min(m,n)}\binom{m}{i}\binom{n}{i}\frac{(il-1)(1-e^{-2il})}{4i^3} \\ &+ \sum_{i=1}^{m}\sum_{j=1,\,i\neq j}^{n}\binom{m}{i}\binom{n}{j} \\ &\times \frac{2ijl - i - j + (i^2e^{-2jl} - j^2e^{-2il})/(i-j)}{8i^2j^2}\Bigg\}, \end{aligned} \tag{22}$$

where the summation terms are included only when the upper limits exceed zero. Note that $\Phi_{m,n}(l) = \Phi_{n,m}(l)$. Despite its complex appearance, eqn (22) is quick and easy to compute.

For lineal descendants that are g generations apart, the increase in the joint allele sharing probability over that for unlinked loci is therefore

$$E(\breve{k}_{1h}\breve{k}_{1i}\breve{k}_{1j}) - k_1^3 = \left(\tfrac{1}{2}\right)^{g-1}\,\Phi_{g-1,\,g-1}(l)$$

and for half-sibs, from eqn (19), it is

$$\begin{aligned} E(\breve{k}_{1h}\breve{k}_{1i}\breve{k}_{1j}) - k_1^3 &= \tfrac{1}{2}[4\Phi_{2,2}(l) - 4\Phi_{2,1}(l) - 4\Phi_{1,2}(l) \\ &\quad + 4\Phi_{1,1}(l) + 2\Phi_{2,0}(l) + 2\Phi_{0,2}(l) \\ &\quad - 2\Phi_{1,0}(l) - 2\Phi_{0,1}(l) + \Phi_{0,0}(l)] \\ &= \tfrac{1}{2}[4\Phi_{2,2}(l) - 8\Phi_{2,1}(l) + 4\Phi_{1,1}(l) \\ &\quad + 4\Phi_{2,0}(l) - 4\Phi_{1,0}(l) + \Phi_{0,0}(l)] \end{aligned}$$

For subsequent generations, e.g. half-cousins, the formulae can be simply extended by methods similar to those used previously for pairs of loci and therefore have the same basic form. These and other results, including those for full sibs and their descendants, are given in Box 2.

For multiple chromosomes that have the same genome content and map length, the skew and variances would be the same for each, and the skewness for whole-genome actual allele sharing would decrease with the square root of the number of chromosomes.

(ii) *Examples*

The magnitude of skew, expressed as the skew coefficient, is illustrated for single chromosomes in Fig. 4 for a wide range of descendants of full sibs and for alternative ancestry, respectively. The magnitude of the skew rises as relationships become smaller, as expected since it is $(1-2k)/\sqrt{[k(1-k)]}$ for single or completely linked loci. Thus, for second cousins, for example, the skew coefficient exceeds 2 even for long chromosomes.

Box 2. *Summary of formulae for skew of genome sharing*

Lineal descendants, where $g=2$ is grandparent–grandoffspring ($k_2=0$)

$$E(\breve{k}_{1h}\breve{k}_{1i}\breve{k}_{1j})-k_1^3=(\tfrac{1}{2})^{g-1}\Phi_{g-1,g-1}(l).$$

Half-sibs and their descendants, where $g=2$ for half-sibs ($k_2=0$)

$$E(\breve{k}_{1h}\breve{k}_{1i}\breve{k}_{1j})-k_1^3=(\tfrac{1}{2})^{g-1}[4\Phi_{g,g,}(l)-8\Phi_{g,g-1}(l)+4\Phi_{g-1,g-1}(l)+4\Phi_{g,g-2}(l)-4\Phi_{g-1,g-2}(l)+\Phi_{g-2,g-2}(l)].$$

Full sibs and their descendants

The actual relationship $\breve{R}$ and also $\breve{k}_1$ for full sibs are symmetrically distributed (Table 1) although the non-central moments are non-zero. The third moment of $\breve{k}_2$ and of $\breve{k}_0$ for full sibs is

$$\begin{aligned}E(\breve{k}_{2h}\breve{k}_{2i}\breve{k}_{2j})-k_2^3=E(\breve{k}_{0h}\breve{k}_{0i}\breve{k}_{0j})-k_0^3=\tfrac{1}{4}[&16\Phi_{4,4}(l)-64\Phi_{4,3}(l)+64\Phi_{4,2}(l)\\&+64\Phi_{3,3}(l)-32\Phi_{4,1}(l)-128\Phi_{3,2}(l)+8\Phi_{4,0}(l)+64\Phi_{3,1}(l)+64\Phi_{2,2}(l)\\&-16\Phi_{3,0}(l)-64\Phi_{2,1}(l)+16\Phi_{2,0}(l)+16\Phi_{1,1}(l)-8\Phi_{1,0}(l)+\Phi_{0,0}(l)].\end{aligned}$$

Uncle–nephew ($g=2$) and descendants ($k_2=0$)

$$\begin{aligned}E(\breve{k}_{1h}\breve{k}_{1i}\breve{k}_{1j})-k_1^3=(\tfrac{1}{2})^{g}[&16\Phi_{g+1,g+1}(l)-48\Phi_{g+1,g}(l)+24\Phi_{g+1,g-1}(l)+40\Phi_{g,g}(l)\\&+4\Phi_{g,g-2}(l)-44\Phi_{g,g-1}(l)+13\Phi_{g-1,g-1}(l)-4\Phi_{g-1,g-2}(l)+\Phi_{g-2,g-2}(l)].\end{aligned}$$

Cousins ($g=3$) and descendants ($k_2=0$)

$$\begin{aligned}E(\breve{k}_{1h}\breve{k}_{1i}\breve{k}_{1j})-k_1^3=(\tfrac{1}{2})^{g}[&8\Phi_{g+1,g+1}(l)-48\Phi_{g+1,g}(l)+56\Phi_{g+1,g-1}(l)+72\Phi_{g,g}(l)-32\Phi_{g+1,g-2}(l)\\&-160\Phi_{g,g-1}(l)+8\Phi_{g+1,g-3}(l)+80\Phi_{g,g-2}(l)+87\Phi_{g-1,g-1}(l)-16\Phi_{g,g-3}(l)\\&-84\Phi_{g-1,g-2}(l)+16\Phi_{g-1,g-3}(l)+20\Phi_{g-2,g-2}(l)-8\Phi_{g-2,g-3}(l)+\Phi_{g-3,g-3}(l)]\ .\end{aligned}$$

4. Variation in actual inbreeding

If an individual's parents are related, it is inbred. At a locus i, the actual inbreeding $\breve{F}_i$ takes values of 0 (alleles not ibd) or 1 (alleles ibd). It has expectation $E(\breve{F}_i)=F$, where F is the pedigree inbreeding, which in turn equals the co-ancestry, $\theta=\frac{1}{2}R$, of its parents. The variance of $\breve{F}_i$ in a population of similarly inbred but independent individuals is $F(1-F)$. Slate *et al*. (2004) analyse the correlation between multi-locus heterozygosity, a function of actual inbreeding, and the pedigree inbreeding, and show how weak this correlation is. Their analysis does not incorporate linkage, however.

For the genome as a whole, the actual inbreeding $\breve{F}$ of an individual is the proportion of its genome which is ibd, with $E(\breve{F})=F$. Linkage affects variation in the actual relationship of individuals with the same pedigree relationship and also therefore increases variation in the actual inbreeding of their offspring. We use an example to show how it can be computed. Individuals **E** and **F** in Fig. 1 are full sibs, and so if they had mated for producing an offspring **X**, the expected inbreeding coefficient of **X** would be 0·25. If **B** is a male, then **M** is a paternal half sib of **X**, **N** is a maternal half sib of **X**, and their offspring **H** and **I** are cousins. The gametes transmitted by **E** to **H** and to **X** have the same random distribution as do those transmitted by **F** to **I** and **X**. Hence, the distribution of $\breve{F}$ of **X** is identical to the distribution of $\breve{k}_1$ of **H** and **I**, who are cousins in this example. From eqn (14) or Box 1 (descendants of full sibs with g=3), $\mathrm{Var}_{\mathrm{FS}}(\breve{F},l)=\mathrm{Var}_{\mathrm{FC}}(\breve{k}_1,l)=8\phi_4(l)-4\phi_3(l)+\frac{3}{2}\phi_2(l)-\frac{1}{2}\phi_1(l)$, which also equals $4\mathrm{Var}_{\mathrm{FC}}(\breve{R},l)$ and $16\mathrm{Var}_{\mathrm{FC}}(\breve{\theta},l)$. Skew coefficients for the actual inbreeding can be obtained similarly.

The arguments do not depend (although the detailed results do) on the relationship among the parents, and can be regarded as a consequence of extending the co-ancestry concept to identity at multiple loci. We are using a quantity, the '*genomic coancestry*', which for a pair of individuals **Y** and **Z** is the proportion of the genome-shared ibd between a random gamete from **Y** and a random gamete from **Z**. Thus, genomic coancestry describes genomes transmitted from individuals, whereas genome sharing (k) describes genomes that are in individuals. Actual inbreeding depends on the genomic coancestry of the two gametes one individual receives; genome sharing and actual relationship depend on the genomic coancestry of the gametes two different individuals

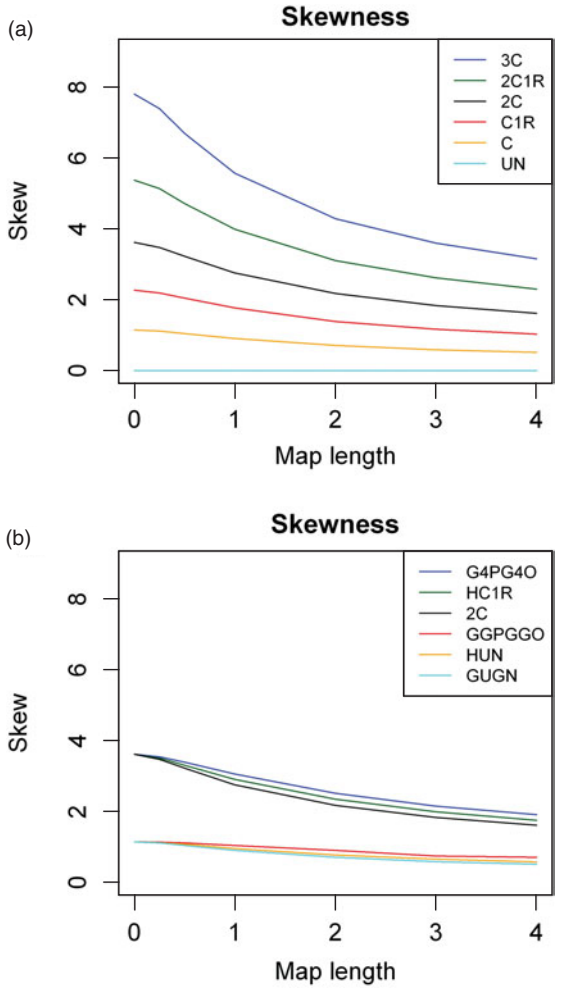

Fig. 4. Skewness of actual relationship (proportion of genome shared) for a single chromosome as a function of map length and relationship for (*a*) descendants of full sibs (as Fig. 2), and (*b*) for different pedigrees for two different degrees of relationships (as Fig. 3). For full sibs and uncle–nephew there is no skew.

receive. For example, the variation of $\breve{F}$ of offspring of cousin matings is the same as that of $\breve{k}_1$ of second cousins, as both are the variance in the genomic coancestry of cousins.

The results for variances, SD, CV and skew of actual relationship given in the Figures and Tables can therefore also be applied directly to actual inbreeding. For example, from Table 2 the SD of $\breve{F}$ of offspring of full sib matings in humans is $2 \times 0{\cdot}0218 = 0{\cdot}0436$ (from item C) and 0·0240 (from item 2C) for offspring of cousins, with the CV of the latter being $0{\cdot}0240/0{\cdot}0625 = 0{\cdot}384$.

The above result applies to the variation in actual inbreeding among a group of unrelated individuals whose parents all have the same pedigree, e.g. are full sibs. In any population there is variation in pedigree inbreeding which also contributes to the total variance in actual inbreeding. The expected variation and distribution of shared segments in any population therefore depend on the population size and mating system, and relevant results for closed populations have been published (Bennett, 1954; Franklin, 1977; Stam, 1980; Weir *et al.*, 1980).

The variation in actual inbreeding can be partitioned into two components, that between families, i.e. the covariance in actual inbreeding of (e.g. full sib) family members, and the variation in actual inbreeding among (e.g. full sib) family members. When we consider just pedigree inbreeding the variance between families is the variance of the co-ancestry from pedigree of the parents, which equals one-quarter of the pedigree relationship of the parents, and there is no variation in pedigree inbreeding within families.

Hence, for full sib matings, for example, $\mathrm{VarB}_{\mathrm{FS}}(\breve{I}, l) = \frac{1}{4}\mathrm{Var}_{\mathrm{FS}}(\breve{R}, l)$. The variance within families can be obtained by difference, and so from the above results for full sib matings,

$$\begin{aligned}\mathrm{VarW}_{\mathrm{FS}}(\breve{I}, l) &= \mathrm{Var}_{\mathrm{FS}}(\breve{I}, l) - \mathrm{VarB}_{\mathrm{FS}}(\breve{I}, l)\\ &= 4\mathrm{Var}_{\mathrm{C}}(\breve{R}, l) - \tfrac{1}{4}\mathrm{Var}_{\mathrm{FS}}(\breve{R}, l).\end{aligned}$$

This can also be regarded as the variance in genomic coancestry of full sibs less the variance in genomic coancestry between their parents.

As an example, using results from Table 2 for the human genome as a whole, $\mathrm{Var}_{\mathrm{FS}}(\breve{I}, L) = 4(0{\cdot}0218)^2 = 0{\cdot}00191$, $\mathrm{VarB}_{\mathrm{FS}}(\breve{I}, L) = (0{\cdot}0392)^2/4 = 0{\cdot}00038$ and $\mathrm{VarW}_{\mathrm{FS}}(\breve{I}, L) = 0{\cdot}00152$, with corresponding SD equal to 0·0436, 0·0196 and 0·0390, respectively. In Table 3, we list relevant relationships and results. It is seen that the variation is substantial and is primarily within families (exclusively within families for selfing and parent–offspring matings of non-inbred individuals). For example, for cousin matings of humans, the mean F is 0·0625 and the SD within families is predicted to be 0·0214.

Estimation of inbreeding depression is usually undertaken by regression of phenotype on pedigree inbreeding. The method can be enhanced by using dense marker data in order to infer the proportion of the offspring genotype that is ibd from the parents and hence actual inbreeding $\breve{F}$ (Slate *et al.*, 2004). By undertaking the analysis within families, confounding environmental effects can be eliminated, with the method being analogous to that of Visscher *et al.* (2006) for estimating heritability within families, but focused on means rather than variances. The design is likely to be most useful for species such as pigs that have large families. Christensen *et al.* (1996) undertook such an analysis, but had only 21 markers available for estimating actual inbreeding (which they refer to as 'realized inbreeding').

5. Discussion

We have shown how to compute the variation and skew in the proportion of genomes shared for diverse

Table 3. *SD of actual inbreeding $\breve{F}$ for a model human genome[a] for matings of relatives*

Relationship of mates	Pedigree F	Relationship-equivalent offspring[b]	$\text{Var}(\breve{F}) \times 10^4$			$\text{SD}(\breve{F})$		
			Betw	Within	Total	Betw	Within	Total
Selfing	0·5	Half sibs	0	30·8	30·8	0	0·0555	0·0555
			0	94·3[c]	94·3[c]	0	0·0971[c]	0·0971[c]
Offspring–parent	0·25	Half uncle–nephew	0	26·2	26·2	0	0·0512	0·0512
Full sibs	0·25	Cousins	3·84	15·2	19·1	0·0196	0·0390	0·0436
Half sibs	0·125	Half cousins	1·92	12·2	14·1	0·0135	0·0350	0·0376
Uncle–niece	0·125	Cousins once removed	1·64	9·92	11·6	0·0128	0·0315	0·0340
Half uncle–niece	0·0625	Half cousins once removed	1·64	5·13	6·76	0·0128	0·0226	0·0260
Cousins	0·0625	Second cousins	1·19	4·57	5·76	0·0109	0·0214	0·0240
Cousins once removed	0·03125	Second cousins once removed	0·72	1·97	2·69	0·0085	0·0140	0·0164

[a] Differing map lengths as in Table 2, except as [c].
[b] Relationship of non-inbred offspring with the same genomic coancestry as the inbred offspring.
[c] For a model maize genome of 10 chromosomes each of 1 M.

kinds of relatives. As theoretical papers have shown previously (Hill, 1993*a*, *b*; Guo, 1995; Visscher, 2009), and anticipated by analyses of junctions and the distribution as a whole, the variance can be high, illustrated most clearly by the coefficient of variation (Fig. 2*b*) and skew (Fig. 4) for increasingly distant relatives.

As the CV is large for single chromosomes each of the average length of those of humans (*c*. 1·6 M) (Fig. 2*b*), exceeding two for second cousins or more distant relatives (Fig. 2*b*), there is substantial overlap in the amount of sharing of quite different pedigree relationship classes. Further, there is substantial positive skew in the distribution over the whole genome for these and more distant relatives, such that individuals with low-pedigree relationship may share much more genome than expected.

In identifying distant relatives in a sample of individuals on which dense SNP data are available, information on potential relationship is available both from estimates of the mean proportion shared and from the variation among chromosomes. That this variation is substantially illustrated by the CVs of actual relationship (Fig. 2*b*), which can greatly exceed unity. Distant relatives are expected to share little or none of the genome of a common ancestor ibd for some chromosomes and a non-negligible amount for others. Indeed, our results for variance in sharing of single chromosomes among pairs of individuals also apply to the variation in sharing among chromosomes of the same length between the same individuals. How best to use such an information has not, in so far as we know, been investigated.

The problem of inferring pedigree relationship from actual relationship (as measured by genome shared) is illustrated in Fig. 5 using the human model genome example. Information on, for example, the distribution of the lengths of shared segments, which will tend to

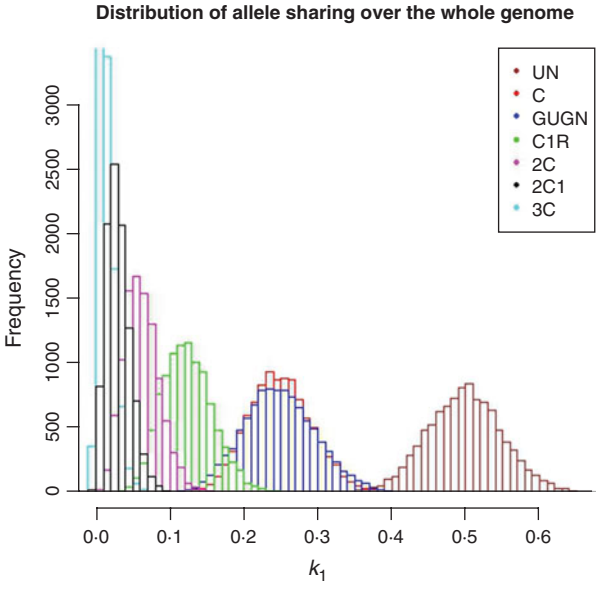

Fig. 5. Distribution of actual genome sharing ($\breve{k}_1$) for samples of 'human' genomes for different degrees of pedigree relationship of descendants of full sibs (as Fig. 2) (10 000 replicates each).

be shorter for distant relatives, also needs to be taken into account, following, for example, the work of Fisher (1954 and earlier), Bennett (1953), Stam (1980) and Thompson (2008) which is based, *inter alia*, on analysis of junctions. Although the distribution of lengths of shared genome that include the end of the chromosome can be computed, there is no general approach that is simple to apply. While it is quite clear that developing methodology using the distributions of chromosome lengths and the numbers of chromosomes for which there is no sharing would be of some interest and potential practical value in establishing pedigree relationship, for example, in forensic situations, such an analysis is beyond the scope of this paper.

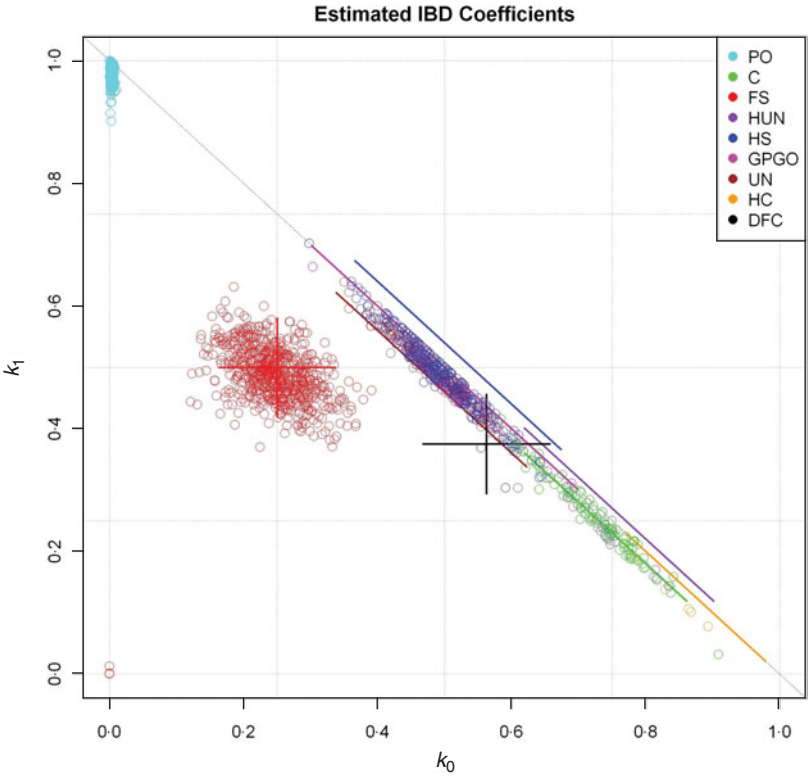

Fig. 6. Estimated ibd coefficients, $\hat{k_0}$ and $\hat{k_1}$, from SNP data for individuals with known pedigree relationship (PO denotes parent-offspring, DFC double first cousins, other symbols as Figs 2 and 3), together with predicted 'error bars' of two SD about expectation. Bars are offset from $k_0+k_1=1$ if $k_2=0$.

Inferring the presence of genes of large effect under selection from shared segments of the genome or for mapping disease genes by comparing allele sharing proportions between affected and unaffected individuals has potential importance, but our results do not give much ground for optimism in its use because the sampling error is so high.

Estimates using dense markers of the variance in actual genome sharing of human full sibs were obtained by Visscher *et al.* (2006, 2007), and, in general, there was good agreement: for example, the observed mean and SD of the proportion of the autosomal genome shared ($\breve{k}_2+\frac{1}{2}\breve{k}_1$), were $0{\cdot}498\pm0{\cdot}036$ compared with expectation $0{\cdot}5\pm0{\cdot}039$, and the corresponding figures for $\breve{k}_2$ were $0{\cdot}248\pm0{\cdot}040$ observed and $0{\cdot}25\pm0{\cdot}044$ expected. The discrepancy was explained by the fact that identical sections could be missed as a limited number of microsatellite markers were used in these studies, averaging 400–600 per individual for the whole genome (Visscher *et al.*, 2006, 2007). We offer further illustration in Fig. 6, using data kindly supplied by Dr M. Marazita. Coefficients of ibd were estimated using SNP data obtained for a whole-genome association analysis of dental caries. Relationship classes were inferred from pedigree information with software developed by Dr Cecelia Laurie and the methods of this paper were used for calculating the SDs of $\breve{k}_0$ and $\breve{k}_1$. For each pair of related individuals in the study (pedigree $R>1/32$), the estimated IBD coefficients ($\hat{k_0}$ and $\hat{k_1}$) were plotted, along with predicted 'error bars' of two SDs each side of the expected values. For display purposes, these bars were offset from the line $k_0+k_1=1$ in the cases for which $k_2=0$. We did not perform any statistical tests for inferred relationships; the error bars reflect only Mendelian sampling and linkage, and the effects of using sample allele frequencies on variation in estimated ibd coefficients will be discussed elsewhere.

The main objective of this paper was to provide general formulae for computing the variance of shared sites. Obviously there are many other avenues to pursue, but these require different techniques.

We are grateful to Peter Visscher for many helpful comments on previous drafts and to Jinliang Wang for a useful suggestion. This work was supported in part by NIH grants R01 GM075091 and HGU0044446, and by the USS. David Crosslin, University of Washington, plotted the figures. Mary L. Marazita, University of Pittsburgh, consented to inclusion of Fig. 6 that displays results from her study of Dental Caries (supported by NIH grants U01-DE018904 and R01-DE014899, and NIH contract HHSN268200782096C to the Center for Inherited Disease Research for genotyping) as part of the GENEVA project (Cornelis *et al.*, 2010). The paper is dedicated to the memory of Piet Stam for his pioneering work in multi-locus ibd.

References

Ball, F. & Stefanov, V. T. (2005). Evaluation of identity-by-descent probabilities for half-sibs on continuous genome. *Mathematical Biosciences* **196**, 215–225.

Bennett, J. H. (1953). Junctions in inbreeding. *Genetica* **26**, 392–406.

Bennett, J. H. (1954). The distribution of heterogeneity upon inbreeding. *Journal of the Royal Statistical Society, Series B* **16**, 88–99.

Bickeboller, H. & Thompson, E. A. (1996*a*). Distribution of genome shared IBD by half-sibs: approximation by the Possion clumping heuristic. *Theoretical Population Biology* **50**, 66–90.

Bickeboller, H. & Thompson, E. A. (1996*b*). The probability distribution of the amount of an individuals's genome surviving to the following generation. *Genetics* **143**, 1043–1049.

Choi, Y., Wijsman, E. & Weir, B. S. (2009). Case-control association testing in the presence of unknown relationships. *Genetic Epidemiology* **33**, 668–678.

Christensen, K., Fredholm, M., Wintero, A. K., Jorgensen, J. N. & Andersen, S. (1996). Joint effect of 21 marker loci and effect of realized inbreeding on growth in pigs. *Animal Science* **62**, 541–546.

Cockerham, C. C. & Weir, B. S. (1983). Variance of actual inbreeding. *Theoretical Population Biology* **23**, 85–109.

Cornelis, M. C., Agrawal, A., Cole, J. W., Hansel, N. H., Barnes, K. C., Beaty, T. H., Bennett, S. N., Bierut, L. J., Boerwinkle, E., Doheny, K. F., Feenstra, B., Feingold, E., Fornage, M., Haiman, C. A., Harris, E. L., Hayes, M. G., Heit, J. A., Hu, F. B., Kang, J. H., Laurie, C. C., Ling, H., Teri, A., Manolio, T. A., Marazita, M. L., Mathias, R. A., Mirel, D. B., Paschall, J., Pasquale, L. R., Pugh, E. W., Rice, J. P., Udren, J., van Dam, R. M., Wang, X., Wiggs, J. L., Williams, K. & Yu, K. (2010). The Gene, Environment Association Studies Consortium (GENEVA): Maximizing the knowledge obtained from GWAS by collaboration across studies of multiple conditions. *Genetic Epidemiology* **34**, 364–372.

Donnelly, K. P. (1983). The probability that related individuals share some section of the genome identical by descent. *Theoretical Population Biology* **23**, 34–64.

Falconer, D. S. and Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics* 4th ed. Harlow, Essex: Longman.

Fisher, R. A. (1954). A fuller theory of 'Junctions' in inbreeding. *Heredity* **8**, 187–197.

Franklin, I. R. (1977). The distribution of the proportion of the genome which is homozygous by descent in inbred individuals. *Theoretical Population Biology* **11**, 60–80.

Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* **136**, 245–257.

Guo, S.-W. (1995). Proportion of genome shared identical by descent by relatives: concept, computation, and applications. *American Journal of Human Genetics* **56**, 1468–1476.

Haldane, J. B. S. (1919). The combination of linkage values, and the calculation of distances between the loci of linked factors. *Journal of Genetics* **8**, 99–309.

Hill, W. G. (1993*a*). Variation in genetic composition in backcrossing programs. *Journal of Heredity* **84**, 212–213.

Hill, W. G. (1993*b*). Variation in genetic identity within kinships. *Heredity* **71**, 652–653.

Kong, X., Murphy, K., Raj, T., He, C., White, P. S. & Matise, T. C. (2004). A combined physical-linkage map of the human genome. *American Journal of Human Genetics* **75**, 1143–1148.

Laurie, C. C., Doheny, K. F., Mirel, D. B., Pugh, E. W., Bierut, L. J., Bhangale, T., Boehm, F., Caporaso, N. E., Edenberg, H. J., Gabriel, S. B., Harris, E. L., Hu, F. B., Jacobs, K. B., Kraft, P., Landi, M. T., Lumley, T., Manolio, T., McHugh, C., Painter, I., Paschall, J., Rice, J. P., Rice, K. M., Zheng, X. & Weir, B. S., for the GENEVA Investigators. (2010). Quality control and quality assurance in genotypic data for genome-wide association studies. *Genetic Epidemiology* **34**, 591–602.

Meuwissen, T. H. E., Hayes, B. J. & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819–1829.

Slate, J., David, P., Dodds, K. G., Veenvliet, B. A., Glass, B. C., Broad, T. E. & McEwan, J. C. (2004). Understanding the relationship between the inbreeding coefficient and multilocus heterozygosity: theoretical expectations and empirical data. *Heredity* **93**, 255–265.

Stam, P. (1980). The distribution of the fraction of the genome identical by descent in finite populations. *Genetical Research* **35**, 131–155.

Stam, P. & Zeven, A. C. (1981). The theoretical proportion of the donor genome in near-isogenic lines of self fertilizers bred by backcrossing. *Euphytica* **30**, 227–238.

Stefanov, V. T. (2000). Distribution of genome shared identical by descent by two individuals in grandparent-type relationship. *Genetics* **156**, 1403–1410.

Stefanov, V. T. (2004). Distribution of the amount of genetic material from a chromosome segment surviving to the following generation. *Journal of Applied Probability* **41**, 345–354.

Thompson, E. A. (2008). The IBD process along four chromosomes. *Theoretical Population Biology* **73**, 369–373.

Visscher, P. M. (2009). Whole genome approaches to quantitative genetics. *Genetica* **136**, 351–358.

Visscher, P. M., Macgregor, S., Benyamin, B., Zhu, G., Gordon, S., Medland, S., Hill, W. G., Hottenga, J.-J., Willemsen, G., Boomsma, D. I., Liu, Y.-Z., Deng, H.-W., Montgomery, G. W. & Martin, N. G. (2007). Genome partitioning of genetic variation for height from 11,214 sibling pairs. *American Journal of Human Genetics* **81**, 1104–1110.

Visscher, P. M., Medland, S. E., Ferreira, M. A. R., Morley, K. I., Zhu, G., Cornes, B. K., Montgomery, G. W. & Martin, N. G. (2006). Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genetics* **2**, e41. doi: 10.1371/journal.pgen.0020041

Weir, B. S., Anderson, A. D. & Hepler, A. B. (2006). Genetic relatedness analysis: modern data and new challenges. *Nature Reviews Genetics* **7**, 771–780.

Weir, B. S., Avery, P. J. & Hill, W. G. (1980). Effect of mating structure on variation in inbreeding. *Theoretical Population Biology* **18**, 396–429.

Weir, B. S., Cardon, L. R., Anderson, A. D., Nielsen, D. M. & Hill, W. G. (2005). Measures of human population structure show heterogeneity among genomic regions. *Genome Research* **15**, 1468–1476.

Wright, S. (1922). Coefficients of inbreeding and relationship. *American Naturalist* **56**, 330–338.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyhot, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W.,

Goddard, M. E. & Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42**, 565–569.

Yu, J. M., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S. & Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* **38**, 203–208.